USING SELF-ORGANIZING MAPS TO MINE CORRELATIONS BETWEEN VOLUNTEERED TICK OBSERVATIONS AND ENVIRONMENTAL DATA

KHITAM JAZI ALMAAITAH

February, 2015

SUPERVISORS:
Dr. R. Zurita-Milla
Ms. I r. P .W.M. Augustijn

USING SELF-ORGANIZING MAPS TO MINE CORRELATIONS BETWEEN VOLUNTEERED TICK OBSERVATIONS AND ENVIRONMENTAL DATA

KHITAM JAZI ALMAAITAH

Enschede, The Netherlands, February, 2015

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geo-informatics

SUPERVISORS:

Dr. R. Zurita-Milla
Ms. Ir. P .W.M. Augustijn

THESIS ASSESSMENT BOARD:

Prof. Dr. M.J. Kraak (Chair)
Ms A. Hofhuis, MSc (External Examiner, RIVM- Epidemiology and Surveillance)
Dr. R. Zurita-Milla (Member)
Ms Ir. P.W.M. Augustijn (Member)
Ms I. Garcia Marti MSc; (Member)

ABSTRACT

Lyme borreliosis (LB) is a major tick-borne disease widely spread in the Northern hemisphere. The incidence of tick bites and Lyme disease cases have been growing in the Netherlands since mid-90s. Moreover, different studies have shown there is a changing of tick abundance in nature. However, higher ticks abundance in nature does not necessarily translated to a higher tick bites occurrence, unless there is strong factors increase the chance of human exposure to tick bites such as human recreation activities. This problem motivates more research, to achieve a detailed understanding of the factors that have an impact on the tick abundance and their consequences on Public Health. Actually; this kind of problems required monitoring of the study area over successive years. Citizens' contribution in collecting observations is considered relatively recent source for gathering spatio-temporal dataset for research.

The increasing use of web technologies enables citizens to participate in the so-called Citizen Science projects, which allow people to participate in collecting observations. Volunteered Geographic Information (VGI) became a decentralized source for collecting spatio-temporal data based on the contribution of people. VGI has been applied to a wide range of disciplines, including tick bite monitoring. Since 2006, Wageningen University (WUR) and the Netherlands National Institute for Public Health and the Environment (RIVM) have been collecting tick bites reported by citizens through the platforms Natuurkalender (NK) and Tekenradar (TR). This dataset contains at present nearly 30.000 observations of tick bites that we have used for this research to study the correlation between the occurrence of tick bites and several environmental variables using data mining techniques. This study also attempts to test whether human volunteers can contribute to science in understanding ticks dynamics in nature.

To achieve the main objective of this research, a comprehensive workflow was implemented to integrate several datasets which were heterogeneous in their nature and sources. The implementation was required three stages: 1) Data preparations and aggregation stage was intended to structure the dataset into space and time matrix, to make it ready for knowledge discovery process. 2) Data transformation, this stage was implemented to construct a similarity measure. The temporal behavior of the tick bites occurrence was selected as a similarity measure between the spatial units. 3) Data mining and correlation analysis. The complete tick bites dataset was used to train self-organizing map (SOM). SOM was used to cluster the spatial units based on their similarity in the temporal behavior of the tick bites occurrence. Finally, the correlation analysis was attempted to study the relationship between the selected environmental variables and the clusters derived from the SOM, to find if any of these explanatory factors has a clear impact on the temporal behavior of the tick bites occurrence and can be used to predict it.

Eventually, the workflow that we implemented for mining tick bites dataset in this research; is adaptable and can be applied to any spatio-temporal dataset. The analyst required to change few parameters to generalize this workflow to any other case study.

Key words: VGI, Spatio-temporal analysis, Data mining, Self-Organizing Map (SOM), Tick bites dynamics, Correlation analysis.

ACKNOWLEDGEMENTS

All the praises and thanks are to Allah for giving me strength, ability, and determination to achieve my MSc degree.

First and foremost, My deep gratitude to my first supervisor Dr.*R.Zuritta Milla*, thank you for your guidance, support, and encouragement during this research period, from you I learned how to try new things without hesitation.

Special thanks and appreciations for my second supervisor Ms. *Ir. P.W.M. Augustijn* for her support, guidance, and kindness. You made me feel home, far away from home Thank you

I would like to thank my advisor Ms *I. Garcia Marti* for her collaboration and explanations during the research period until the last moment. Thank you

I would like to thank The Netherlands Fellowship Program (NFP) for sponsoring this MSc programme.

I would like to thank all of my friends and classmates, it was my pleasure to meet you all thank you for sharing knowledge, experiences and moments. Special thanks for *Sana, Yolla, Eskedar* and *Maral* for surrounding me like a family.

My success is dedicated to my father and to my mother's soul who were my first teachers in this life. From you I learned how to go in the direction of my ambitions. Special thanks to my sisters and brothers for their prayers and support specially my sister *Mai.*

My deepest gratitude to my great husband *Samer,* this won't be possible without your support. You have demonstrated that a great man could be behind a successful woman. Thank you for being an Arab distinctive man.

I would like to thank my daughter *Judi* for being patient and mature enough during this period. I would apologize for every moment you need me and I was not around. It was hard experience but it made us stronger. One day you will be able to read this and I'm sure you will be proud of your parents.

Khitam Jazi Al-Maaitah,
Enschede,
The Netherlands, February 2015

TABLE OF CONTENTS

1. INTRODUCTION 1
   1.1. Motivation and problem statement 1
   1.2. Research identification 2
   1.3. Project setup 4
2. LITERATURE REVIEW 7
   2.1. Introduction: 7
   2.2. General information about Tick ecology: 7
   2.3. Environmental factors that control tick abundance 9
   2.4. Volunteered geographic information 11
   2.5. Review of methods for correlation detection : 12
   2.6. The Self - Organizing Map (SOM) 13
3. MATERIAL AND METHODS 19
   3.1. Data 19
   3.2. Methods 24
4. RESULT AND DISCUSSIONS 32
   4.1. Data preparation and aggregation 32
   4.2. Data transformation result 39
   4.3. Data maining and correlation analysis 39
5. CONCLUSIONS AND RECOMMENDATIONS 52
   5.1. Conclusions 52
   5.2. Recommendations 54

   LIST OF REFERNCES 56

   APPENDICES 64

LIST OF FIGURES

Figure 1: Factors have an influence on tick-borne disease 3
Figure 2: Research method 5
Figure 3: Ixodid tick life cycle. 8
Figure 4: Geographic distribution of erythema migrans in the Netherlands 9
Figure 5: The typical structure of the SOM. 14
Figure 6: Different shape of the grid a) hexagonal lattice b) Rectangular lattice.. 15
Figure 7: a) U-matrix and b) labels obtained by SOM for Iris dataset 16
Figure 8: Component planes of SOM trained using digester data. 16
Figure 9: Schools regions in the Netherlands. 21
Figure 10: The Developed workflow for research implementation. 25
Figure 11: The over view of data preparation and aggregation stage. 26
Figure 12: Data structure in space and time matrix. 27
Figure 13: Temporal behavior of tick bites occurrence in an arbitrary selected spatial unit (x) 29
Figure 14 : Overview of the VGI data quality check 32
Figure 15: Number of the Volunteers according to their age. 33
Figure 16: Boxplots for the reported tick bites per year. 35
Figure 17: Histograms of the weekly number of reported tick bites per year(2006-2009) 36
Figure 18: Histograms of the weekly number of reported tick bites per year (2010-2013) 37
Figure 19: Data aggregated based on several aggregation levels. 38
Figure 20: Sample of the final structure of ticks bites dataset. 39
Figure 21: SOM output plots 40
Figure 22: SOM clusters 41
Figure 23: Average of de normalized codebook vector 42
Figure 24: Percentage of the input data per cluster 43
Figure 25: Projection of SOM clusters into the geographic space (2006-2009) 44
Figure 26: Projection of SOM clusters into the geographic space (2010-2013) 45
Figure 27: Average of LULC percentage per cluster 47
Figure 28: Average of forest mean patch area per cluster 47
Figure 29: Average weekly temperature per cluster (Year2007) 48
Figure 30: Average weekly temperature per cluster (Year 2009) 49
Figure 31: Number of spatial units that have autumn school vacation per cluster 49
Figure 32: Visual inspection of two spatial units 51

LIST OF TABLES

Table (1): SOM input data..... 14
Table (2): Tick bites dataset attributes..... 19
Table (3): Selected environmental variables..... 20
Table (4): Sample of school's vacations dataset for year (2012/2013)..... 22
Table (5): Environmental data preparation steps in the research..... 23
Table (6): Sample data of an arbitrary selected spatial unit..... 28
Table (7): Calculations of the temporal behavior..... 29
Table (8):Number of reported tick bites (yearly)..... 32
Table (9): Percentage of the input data per cluster (2006-2013)..... 42

LIST OF ABBREVATION

VGI	Volunteered Geographic Information
SOM	Self-Organizing Map
EDA	Exploratory Data Analysis
NDVI	Normalized Difference Vegetation Index
LULC	Land Use Land Cover
OSL	Ordinary Least Square
GWR	Geographically Weighted Regression
WOY	Week of the Year

1. INTRODUCTION

1.1. Motivation and problem statement

The rapid emergence of new data-gathering technologies has facilitated and increased citizen contributions in data-gathering process. In particular, Volunteered Geographic Information (VGI) is considered as part of a web phenomenon called User-Generated Content (UGC) (Goodchild, 2007). The evolution of VGI data was a consequence of the rapid development in W2.0 era applications and technologies, which have enabled users to generate content with reference in space and time. As a result, the volume of VGI spatio-temporal dataset is steadily growing.

Using VGI data as a data source for research has become common regardless of problems related to its quality. First, VGI data is collected by volunteers who are contributors from the community; they are non-experts and untrained. Second, their contribution is not planned and their motivation to contribute towards the data collection process is unknown, even their profile is unknown. In fact, the volunteer's characters have an impact on the quality of VGI data (Karimi, 2014). Third, most of the VGI datasets have a massive volume which made it challenging to process and analyse. Nevertheless, researchers and scientists used VGI as a data source in several disciplines. As long as it is offering big datasets collected at almost no cost and significantly reduced the time for data collection.

In fact, an interesting knowledge is usually embedded in such a spatio-temporal datasets. The problem is how to make sense of geographic data collected by volunteers in order to derive a scientific knowledge and further understanding of the physical phenomenon occurring in space and time surrounding us in this world. Spatio-temporal data mining is a popular research area, capable to discover knowledge in database. The initial step toward discovering knowledge in spatio-temporal data is the development of comprehensive analytic workflow to extract knowledge embedded in spatio-temporal data. This process usually includes initial understanding about the behavior of the phenomenon in nature. Followed by a complex data pre-processing stage; transformations may be required besides applying a proper data mining technique (Bogorny & Shekhar, 2010). The importance of data mining is increasing in different disciplines, and its applications can be found in several domains such as Meteorology, transportation, forestry and ecology (Rao et al., 2012). Moreover, data mining has been applied in studying animal and plant's phenology.

Environmental variables influence the abundance and distribution of animals and plants' species in nature. These distributions vary in nature through time. Ecological research commonly addresses the correlation between the relevant environmental variables and certain animal or plants species over a time span. This kind of studies may lead to better understanding and contribute to solving real-life problems. Specifically, phenology is defined as: study of periodic life cycle events in plants and animal's species. In addition, it studies the impact of climate and environmental factors on these phenological events (Betancourt et al., 2007). As it is known, environmental variables vary through space and time. For this reason, in phenological studies it is necessary to study the phenomenon over a spatial extent a long several successive years. This type of scientific research requires combining several of heterogeneous datasets. For instance, combining spatio-temporal data collected by volunteers (reporting events in nature) with relevant

environmental variables' datasets. The heterogeneity of the spatial data types and sources makes data processing a complex and challenging task.

The growing interest in monitoring and understanding the effect of environmental variables on plants and animals' species motivates the contribution of volunteers in phenological research. For example, In The Netherlands, the National Institute for Public Health and the Environment (RIVM) together with Wageningen University lunched "Tekenradar.nl" in 2012, as a monitoring platform for ticks (Tekenradar.nl, 2014). The project aims to increase the knowledge on how often and under which conditions a tick bite leads to Lyme disease (Doelen Tekenradar.nl, 2014). This research is motivated towards the study of the correlation between tick bites occurrence and environmental variables using data mining techniques. For this purpose, VGI datasets together with weather data, and land cover will be used.

1.2. Research identification

1.2.1. Research problem background

Lyme borreliosis (LB) is a major vector borne disease widely spread in the Northern hemisphere (Mulder et al., 2013), and the most prevalent according to the World Health Organization, (2014). This disease can be transmitted to humans through the bite of a tick carrying a bacterium called *Borrelia burgdorferi*. Ticks need blood at each stage of their life cycle to survive and reproduce. Ticks wait on the grass or leaves for a host (human or animal) to bush the vegetation and then climb onto the host to start feeding. Otherwise the tick dies (Life cycle of Hard Ticks that Spread Disease, 2014). Throughout the different species of ticks, *Ixodes ricinus* is considered as a main vector of *Borrelia burgdorferi* and can convey Lyme disease to humans(Barrios et al., 2012).

Recently, several studies have shown a dramatic increase in the number of *Lyme Borreliosis* incidence in the Netherlands. For instance, Den Boon et al. (2004) found that the number of tick bites doubled in the period between 1994 to 2001. In addition, a study carried out by Hofhuis et al. (2006) shows that there is significant increase in the number of general practitioners (GP) consultation for tick bites, and the number of Lyme disease incidences admitted to hospitals in the Netherlands. The dramatic increase in the number of tick bites indicates an increase of tick abundance and distribution in nature.

Several factors have an influence on tick abundance and tick-borne diseases as shown in Figure (1). In general, a wide distribution of ticks increases the chance of tick bites. In addition, host availability and density have played an important role in this context: On one hand, ticks' movement depends on the mobility of their hosts, which affects their geographic distribution. On the other hand, hosts have a role in maintaining and transmitting disease pathogens. For instance, some small mammals such as white-footed mouse are considered as a main reservoir of a bacterium which causes tick-borne disease (Estrada-Peña and de la Fuente, 2014, Tran and Waller, 2013). Furthermore, human interaction with the environment has an effect on the increase in the number of tick-borne disease cases. Particularly, their close contact with the vegetation in risky areas leads to an increased chance to tick-bite exposure. Moreover, climate variables and land cover have an impact on tick abundance and their activity (Estrada-Peña and de la Fuente, 2014).

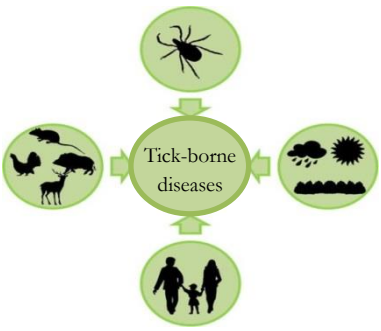


Figure 1: Factors have an influence on tick-borne disease

Nowadays, there is more concern regarding the impact of tick-borne diseases on public health in The Netherlands. Actually, the complex interaction between the ticks with their host, environment and humans (Pfäffle et al., 2013), make it difficult to limit or control tick-borne diseases such as Lyme disease (Dantas-Torres et al., 2012). This problem motivates more researches, in order to achieve a detailed understanding of the factors that have an impact on the ticks' abundance and their consequences. In this research we attempt to integrate several environmental datasets together with VGI dataset to study the impact of environmental variables on tick bites occurrence.

1.2.2. Research objectives

This research aims to study the correlation between the occurrence of tick bites and several environmental variables in the Netherlands based on volunteered dataset and data mining techniques. Achieving the main objective of the research requires fulfilling the following sub-objectives:

- Identify from literature the key environmental variables affecting tick abundance and activity.
- Design a computational workflow to pre-process relevant spatio-temporal data to confirm the consistency of the input data in space and time.
- Implement a Self-organizing map (SOM) to analyse the occurrence of tick bites.
- Visualize and interpret the result of the analysis.

1.2.3. Research questions

Based on the sub-objectives the following research questions supposed to be answered:

- What are the key environmental variables that affect the tick abundance addressed in literature?
- What aggregation level is suitable for the analysis of the tick bites?
- Which exploratory techniques can highlight the period with a higher occurrence tick bites?
- How to handle the problem of VGI data quality?
- What are the most appropriate parameters for the SOM?
- Which environmental variables have more impact on the tick bites occurrence?

1.2.4. Innovation

The innovation in this research comes from:

- Previously, several studies have addressed the impact of the environment on the tick dynamics based on collecting ticks samples from nature. This research implemented based on volunteered geographic information to study the correlation between environmental variables and the occurrence of tick bites is considered new. Here, in this research we testing whether human volunteers can contribute to science by helping in understanding tick dynamics.
- Design a workflow to pre-process heterogonous spatio-temporal datasets which are challenging to integrate.

1.3. Project setup

1.3.1. Research method

The following steps were applied to achieve the objectives and answering the questions of this research:

- Literature review: in the initial stage of the research, several previous studies were reviewed. On one hand, knowledge from literature was required to know which environmental variables affect the tick's abundance and activity. On the other hand, studies about data mining and spatio-temporal analysis are reviewed to know how the previous studies applied data mining techniques to study the correlation between the variables.

- Data preparation and aggregation: In general spatio-temporal volunteered datasets are challenging to combine. In this research, datasets which involved in the analysis have heterogeneous nature. All datasets should be consistent in spatial and temporal scale. During this stage, aggregation was applied to confirm the consistency of the input data. In addition data exploration was done to the VGI dataset to highlight the period that has a significantly high number of tick bites.

- Data transformation: there was significant difference between the numbers of observations in tick bites dataset from year to year. To avoid this difference data transformation was applied.

- Data mining and correlation analysis: The analysis stage aimed to mine the tick bites dataset using Self-Organizing map, which was used to cluster the spatial units according to their similarity. Further, the correlation between the clusters derived from the SOM and the environmental variables such as temperature, LULC percentage, forest fragmentation, and school vacation data were analysed. To study the influence of these variables on tick bites occurrence, statistical calculations were performed; Figure (2) illustrates the proposed methodology for the research.

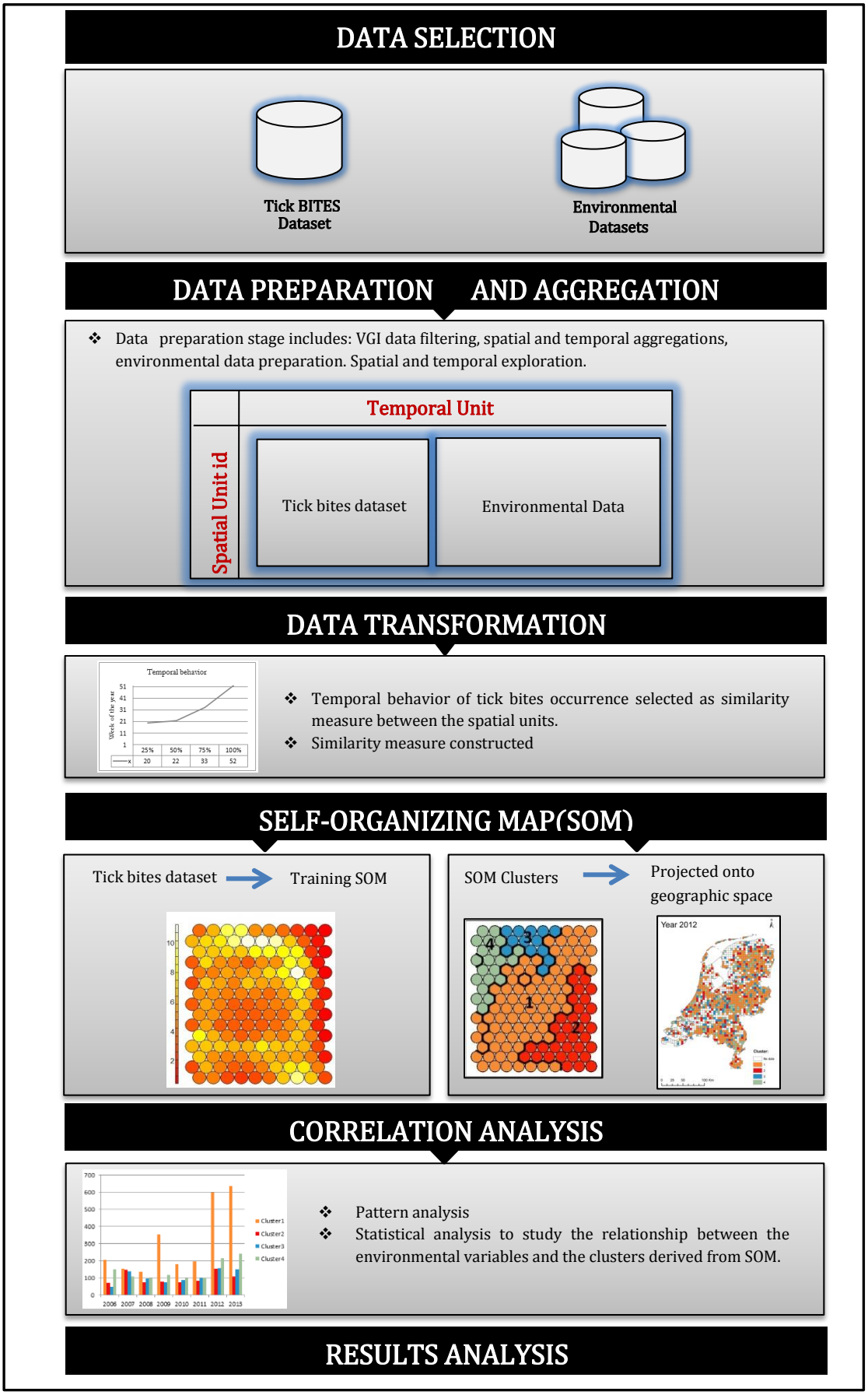


Figure 2: Research method.

1.3.2. Thesis structure

This thesis consists of five chapters. *Chapter 1* introduces the research problem, the motivation, the research objectives and the research questions which were addressed in this study. All literatures and previous studies that were reviewed to understand the theoretical background of this research are described in *chapter2*. This includes all relevant facts about the selected method and background about ticks and environmental variables that have major role on its distribution. *Chapter 3* provides a description of the datasets which are used in the analysis and the methods were applied in this research. *Chapter 4* presents and discusses the results of implementing the analysis. Finally, *Chapter5* provides the conclusions and answers to the research questions in addition to recommendations for future work.

2. LITERATURE REVIEW

2.1. Introduction:

The content of this chapter provides an overview of the most important information and methods regarding this research. The first chapter starts with a general description about tick ecology and a brief explanation of their life cycle. The second section discusses the key environmental factors that control ticks abundance and play a major role in its distribution, while the types of data which are previously used in tick researches are discussed in the third section. In addition, section four provides an essential description of self-organizing map (SOM). Finally, the usage of SOM as a tool for correlation detection is explained in the last section, which is the method applied in this research.

2.2. General information about Tick ecology:

Ticks are blood-sucking arthropods belonging to the subclass *Acari* that parasitize hosts (Sobrino et al., 2012). There are around 850 tick species in nature divided into two main families, *Argasidae* (soft ticks) and *Ixodidae* (hard ticks)(Parola and Raoult, 2001; Estrada-Peña and de la Fuente, 2014). Despite the fact that ticks look like an insect, one can distinguish them from insects by a few significant differences. For instance, the tick's body is segmented into two parts while the insect's body has three segments. Ticks do not have wings and can't fly, so they; mainly depend on their host mobility to move. Furthermore, ticks can be identified by the number of legs, which are different according to their life stage. Particularly adult ticks have eight legs while larvae has only six legs (Life cycle of Hard Ticks that Spread Disease, 2014).

Typically, a tick passes through three stages in the life cycle after hatching from eggs, which are larvae, nymph, and adult (Sandberg et al., 1992). Ticks need a blood meal to survive and molt from one stage to the next stage. Therefore, ticks seek a host (e.g. mammals, birds and reptiles) to attach to and start feeding up to several days. After feeding they detach and drop on the ground to digest the blood meal and to molt to the next stage. In general, Ixodid ticks need between two and three years to complete their life cycle. The length of their life span depends on the suitability of the environmental conditions. Figure (3) illustrates the life cycle of Ixodid ticks: after hatching from eggs, larvae start questing on the vegetation waiting for a host to attach to and start feeding (Figure3 – A). Then it drops to the ground to digest the blood meal and molts to the next life stage which is the nymph (Figure3 – B). Ticks need a blood meal in each life stage to survive. Therefore, nymph will seek again for the second host and attach to it for the following blood meal (Figure3 – C). Then the nymph drops off another time and moults into an adult, which is the last stage of the tick life cycle (Figure3 – D)(Estrada-Peña and de la Fuente, 2014). The resulting adult attaches to the third host to feed and mate and finally the females drop of to lay thousands of eggs.

Ticks can occupy three hosts during their life cycle. In nature ticks are questing on different heights, which have correlation with the host size, for example, larvae questing on low growing vegetation and can be found on hosts of all size (deer, rodents or rabbits), while nymphs and adults are commonly on a large-size host only. This is because they are questing on heights above the areas where rodents, rabbits (smaller hosts) are running (Randolph, 2013).

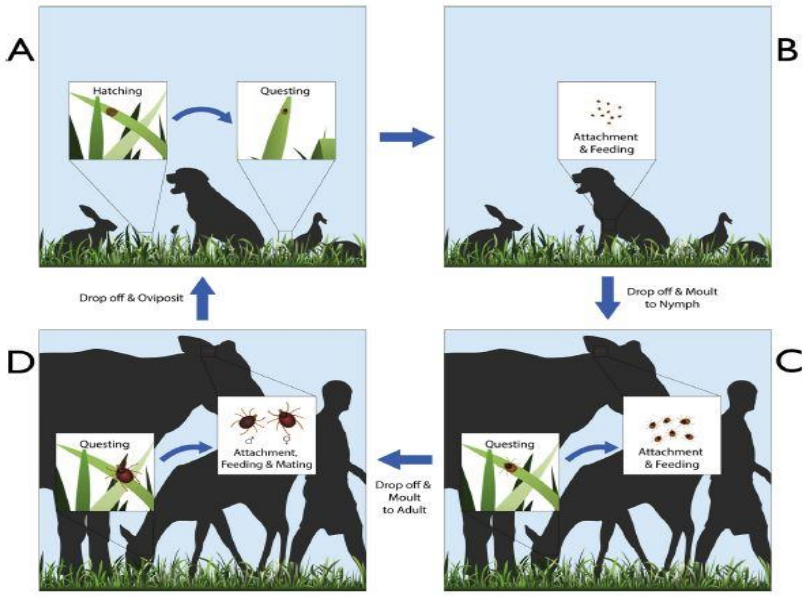


Figure 3: Ixodid tick life cycle. Source: Estrada-Peña and de la Fuente, 2014.p

Ticks are able to carry and transmit a wide variety of pathogens to their host (human or animal) (Fritz, 2009; Zhang et al., 2012), such as viruses, bacteria protozoa and helminthes(Baneth, 2014). Currently, ticks are considered the second most important vector after mosquitos in pathogens transmission to humans (Parola and Raoult, 2001; Vu Hai et al., 2014). As discussed before, ticks feed by biting hosts and, as a consequence of the blood-sucking process, there is potential for the tick to carry the pathogens from an infected to a healthy host. This is how they spread diseases. Usually, ticks wait on the grass or leaves for a host (human or animal) to bush the vegetation and then climb onto them. It takes from ten minutes to two hours for ticks to find a feeding spot on the host skin (How ticks spread disease, 2014). And then ticks anchor their mouth to the host skin to start feeding. Baneth (2014), mentioned that, most tick-borne infections are transmitted to humans through tick's saliva, which contains compounds with anti-hemostatic and anesthetic properties, so human do not notice the tick bite (Coons and Rothschild, 2004). Therefore, putting small amount of saliva on the feed spot helps the tick in sucking adequate amount of blood meal and makes the tick bite painless to their hosts. Typically, transmitting the pathogen of the disease takes 24 – 48 hours after the tick attaches to the human. Therefore, tick removal within 24 hours reduces the risk of passing tick-borne disease (How do humans get Lyme disease, 2014).

Nowadays, tick-borne diseases form a public health concern in Europe. In general, *Ixodid ticks* (hard ticks) are the most common tick species in that region (Petney et al., 2012; Barrios et al., 2013). Particularly, Ixodes may act as living carrier of pathogens, such as, bacteria *Borrelia burgdorferi*, *Rickettsia spp.*, *Babesia spp.* and tick-borne encephalitis virus (TBEV) (Sprong et al., 2014). Among all kinds of tick-borne diseases, Lyme borreliosis is a major vector borne disease widely spread in the Northern hemisphere and the most prevalent vector-borne one (Mulder et al., 2013; WHO, 2014). This disease is caused by bacteria *Borrelia burgdorferi*. Sprong et al. (2014) mentioned that at least 65000 Lyme disease incidences are reported annually in Europe. In particular, several studies have shown a dramatic increase in the number of Lyme Borreliosis incidence in the Netherlands. For instance, Hofhuis et al. (2006) have shown that there is a significant increase in the number of general practitioners (GP) consultation, and the number of Lyme disease incidences. Boon Den et al. (2004) found that the number of tick bites doubled in the period between 1994 and 2001. Figure (4) illustrates the change of the geographic distribution of erythema migrans (which is the early stage of the Lyme disease).

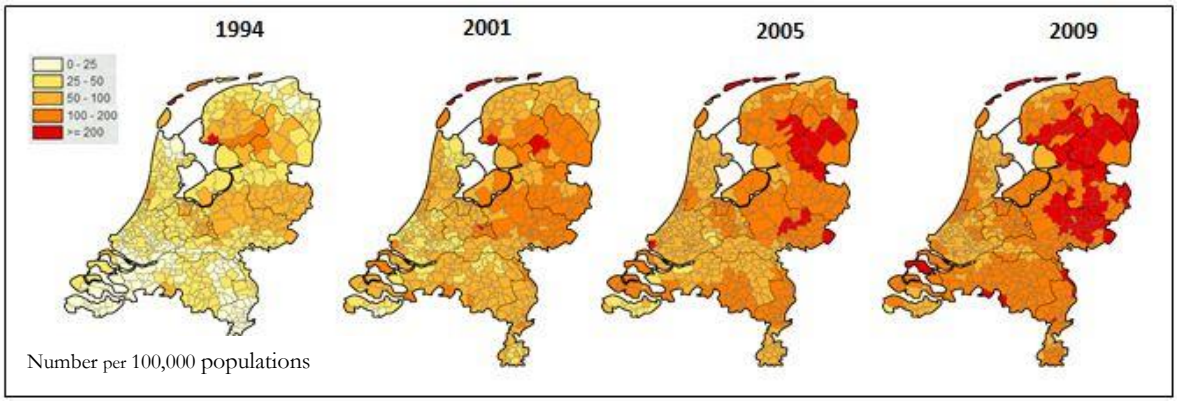


Figure 4: Geographic distribution of erythema migrans in the Netherlands.Source: www.Tekenradar.nl.

Nowadays, there is more concern regarding the impact of tick-borne diseases on public health than in the past. No doubt that preventing tick bites leads to preventing tick borne disease but, the complex interaction between ticks, host, environment and humans (Pfäffle et al., 2013), makes this challenging. In nature ticks are not evenly distributed, they are more abundant in areas that provide suitable conditions for their survival. Typically, ticks have a wide distribution in deciduous woodlands and mixed forests (Agustín Estrada-Peña, 2001; Schwarz et al, 2009). However, ticks can also be found in home gardens in urban areas (Mulder et al., 2013). Outdoor activities such as hunting, camping, mushroom collecting, and farming increase the exposure to ticks. The proper ways to avoid tick bites in these cases are: wearing long sleeves have trousers tucked in the boots and use an insect repellent on uncovered skin. However, inspection of the entire body and pets after outdoor activities is strongly recommended because the rapid tick removal within 24 hours reduces the disease transmission likelihood (Fix, 2006).

2.3. Environmental factors that control tick abundance.

Ticks are present in nature and affected by the surrounding environment, Several environmental factors affect tick abundance and control the dynamic of its distribution such as : climatic factors, land cover, landscape and host availability (human or animal) (Hubálek et al., 2006; Rosà and Pugliese, 2007). This section briefly discusses the key environmental variables that are considered as key factors affecting ticks.

- **Climatic factors**: tick distribution can vary from year to year according to the climate conditions which have a strong influence on tick distribution (Vu Hai et al., 2014). In particular, the temperature and humidity have a major role in regulating tick activity and development (Boyard et al., 2007).In their paper, Sprong et al., (2012) mentioned that nymph and adult ticks become active and start seeking for a blood meal when the daily maximum temperature exceeds 7°C. While questing on top of the vegetation, a tick may lose its water content, actually the water balance in tick is influenced by the relative humidity and water saturation in the air (Estrada-Peña et al., 2013).Several researches studied the correlation between the tick abundance and the climate variables. For instance, Li et al. (2012) have included climatic and environmental factors such as, temperature, humidity, wind speed and forest fragmentation in a multi-level analysis, in order to study the spatio-temporal dynamics of *Ixodes ricinus* in Belgium. The results show, negative relationship between wind speed, relative humidity, and daily abundance of questing nymphs. Bennet et al. (2006) included mean temperature, relative humidity, and precipitation in a Poisson's regression model and found that humidity affects the ticks' abundance. Their results confirmed previous studies, which have shown that the ticks dehydrate and cannot seek for host

when humidity is below 86%.While(Bennet et al., 2006, Subak, 2003) show that there is a positive correlation between humidity and Lyme disease incidence. Humidity has been included in a different way by Barrios et al. (2012), who used the Normalized Difference Water Index (NDWI) as an indicator of moisture in vegetation. Their study led to consider NDWI as a suitable indicator of Lyme incidence and spread of ticks. No doubt that relative humidity was affected by pattern of precipitation at the regional scale However, there is no universal association between the relative humidity and precipitation(Estrada-Peña, A. et al., 2013). This research includes the minimum and maximum mean temperature and as it has an impact on tick abundance and human recreational behavior.

- **Land cover**: As mentioned before ticks are usually found in vegetated areas such as forest. Most of the researches in North America and Europe show that the forests are suitable habitat for ticks and determine their spatial dynamics (Li et al., 2012). The humid and cooler conditions make the forests as more favourable environment for ticks rather than green areas such as pastures (Tack et al., 2012). Here, the vegetation plays two roles in this context; firstly it provides questing sites for ticks while seeking for a host. Secondly, it maintains the microclimate conditions which are immediately surrounding the ticks and required for their survival and developments. Moreover, the spatial distribution of ticks and tick-borne diseases like Lyme disease has been linked with the spatial distribution of vegetated areas (Barrios et al.,2013). The Normalized Difference Vegetation Index (NDVI) has been used by Kalluri et al. (2007) who linked the NDVI with tick abundance and tick-borne disease transmission.

- **Landscape structure:** Researches address the impact of forest fragmentation on tick abundance. Forest fragmentation is a phenomenon occur when continuous forest is fragmented in to several small patches either by natural process (e.g. fire or climate) or human activities (e.g. urbanization, cleaning for agriculture). This, leads to biological and physical changes in the forest environment (Jha et al., 2005). Forest fragmentation is considered one of the factors that have an impact on tick abundance. Particularly, it has potential consequences for the number of Lyme disease incidences. Tran and Waller (2013) linked Lyme disease incidences with landscape and climate variables to feed a negative binomial regression model. The result of this study showed that more fragmentation in deciduous forests has a positive relationship with Lyme disease cases in the United States. This result can be justified by the knowledge that forest fragmentation leads to reduce the density of mammalian species, and increasing the density of white-footed mice. In fact, these small rodents are considered as a reservoir of Lyme bacterium (*Borrelia burgdorferi*) and have a major role in transmission the infection to the ticks (Tran & Waller, 2013). On the other hand, hosts have a potential role in increasing the risk of the tick spatial distribution, as long as the ticks have limited ability to move over long distance in nature (Kalluri et al., 2007). In this research mean patch area is used as an indicator of forest fragmentation.

  Recent studies have addressed the association between tick phenology and altitude. Particularly, the effect of altitude on tick (*I. ricinus*) abundance has been tested by Gilbert (2010) in Scotland. The study carried out in nine sites and the results have shown strong negative association of tick abundance with the altitude. However, this research is not intended to include altitude, due to the uniform elevation in the Netherlands.

2.4. Volunteered geographic information

This section provides a brief description of several types of data that have been used in research on ticks previously. Besides, an overview on volunteered geographic (VGI) data as a data source that will be used for this research.

The data collection stage is an initial and critical step for all the studies on ticks ecology, abundance, and the dynamics of tick-borne diseases. There are different procedures to collect data about ticks. Blanket dragging (or flagging) is a commonly used procedure for collecting questing ticks (i.e. the ones that are waiting on vegetation for a host to pass by). Collecting tick samples by dragging is applied by using a piece of fabric (approximately 1 m^2) which is passed over vegetated areas. Dragging is implemented based on some rules to collect accurate samples for reliable results. For instance, conduct the sampling over homogenous patches of vegetation within a defined time interval. Data collected by dragging: on one hand, can be used to estimate tick activity rate. On the other hand, it can be compared with data from different time period and associated with data about abiotic factors (e.g., climatic variables) to study the impact of the environmental variables on ticks' activity and distribution (Estrada-Peña et al., 2013).

Other researchers were studied the dynamic of tick-borne disease based on collecting data about tick bites is using surveillance system, which is based on information collected by tick disease specialist and general practitioners (GP). This type of data has been used before to understand the dynamic of tick-borne disease and estimate the risk of these diseases on humans. Therefore, a questionnaire survey and Hospital databases can be used to record tick-borne disease cases reported (Vu Hai et al., 2014).

Recently, another alternative is to relay on volunteers. Volunteered geographic information (VGI) becomes a novel source for collecting spatio-temporal data. Several studies were used VGI data in different disciplines such as phenology. In phenological studies, volunteers were involved to collect information about the phenological phases of the species. In this research VGI dataset was used; here, the volunteers were reporting tick bites incidence. This study attempted to test whether human volunteers can contribute to the science in understanding ticks dynamics in nature

Volunteered geographic information (VGI) is considered relatively recent but quite prominent phenomenon where private citizens' contribute to create geographic information (Goodchild, 2007). In his article, Goodchild (2009) considered VGI as a part of a web phenomenon called user-generated content, and he introduced the concept of "citizen science". The contribution of the volunteers (citizen scientists) in collecting the data is not a new idea, it was started before more than a century in ornithology studies as reported in (Wiersma, 2010). Currently, the opportunity of creating and sharing VGI data has engaged many individuals' interest. Goodchild, list the reasons behind citizens' contribution in data gathering by: self-promotion, personal satisfaction, and altruism.

Nowadays, the usage of VGI data becomes more popular in different disciplines and it has proven its efficiency in offering an alternative way for geographic data acquisition for research. Because, individuals' contribution in acquiring detailed geographic information is cheaper than any other alternatives. But, there is general lack of the VGI data quality (Goodchild and Li, 2012). Generally, the citizens are not trained and their contribution is voluntary. Therefore, the result might have sufficient accuracy but this might also not be the case. Nevertheless, VGI forms a dramatic innovation influencing several disciplines of geography. The emergence of VGI can be justified on one hand by, the rapid advances of web 2.0 technologies. On the other hand, the availability and wide spread use of smart devices which are providing

positioning sensors such as smart phones. These technologies enable the contribution of the community in creating, managing, and sharing geographic data via the World Wide Web.

According to the US spatial data transfer there are five fundamental standards for VGI quality as mentioned in (Goodchild & Li, 2012b) which are : positional and attribute accuracy, completeness, logical consistency and data lineage. In fact several studies have been addressed the quality of VGI data and attempt to evaluate its accuracy. For instance, Haklay, (2010) has been evaluated the accuracy of OpenStreetMap by comparing it data with reference data form national mapping agency and the result he found approximately 6 m positional displacement. Regardless, the lack in VGI quality and it does not follow the scientific rules of sampling but it may be beneficial in the initial exploration stage and guide researcher in hypothesis generation (Goodchild & Li, 2012b).

Involving citizens in collecting data about tick bites started in Europe. To be precise, in the Netherlands, the National Institute for Public Health and the Environment (RIVM), and Wageningen University lunched "Tekenradar.nl" in 2012 (Tekenradar.nl (Tick radar), 2014). The project aims to increase the knowledge on how often and under which conditions a tick bite leads to Lyme disease (Doelen Tekenradar.nl, 2014). The project involves the community in collecting volunteered observations in order to implement researches on Lyme disease and tick bites. This website (Tekenradar.nl), allows for citizens who have been bitten by ticks to report a tick bite or an erythema migrans case. Tekenradar is a spin-off of an older volunteered based program called Natuurkalender which started collected tick bites data in 2006. In this research a VGI dataset collected from Tekenradar and Natuurkalender used and related with relevant environmental datasets to study the impact of several environmental variables on tick bite occurrence.

2.5. Review of methods for correlation detection :

Previously, several researchers have addressed the correlation between environmental factors and species distribution in nature. In the case of tick phenological studies, the most commonly used approaches were statistical methods, which are usually based on an assumption that the environmental variables have an impact on: tick development rate and abundance, pathogen transmission, human exposure to tick bites and habitat of the ticks and their host. For example, Wimberly et al., (2008) used a hierarchical Bayesian modelling approach to study presence/absence of pathogen at the country level in the south-eastern United States, they justified their choice for this approach base on the fact that it allowed to study the spatial auto correlation and environmental correlation between climate, land cover, host population and risk of tick-borne disease. Moreover, Brownstein et al. (2003) developed a spatial logistics model to predict the distribution of Ixodes scapularis at country level in 48 states belonging to the United States. The result shows that the temperature and vapor pressure have a major contribution to population distribution. Nevertheless, Estrada-Peña et al. (2006) state that the statistical prediction models which are built based on small regions has limitations and may not be applied for another areas which may lead to lack of prediction.

Typically, the spatial analysis aims to study the relationships between the explanatory variables. This can be identified by statistical calculations in order to estimate model parameters based on observations from the study area. Statistical methods assumed that the parameters are constant over all locations. Although there might be differences in the relationships between the variables over space(Brunsdon et al., 1996). In the case of spatio-temporal data analysis, the data consists of series of spatial data overtime. The empirical application that deals with such a type of spatio-temporal data are assuming that the temporal dimension can be neglected as a result the estimation is based on the spatial dimension only (Dube and Legros, 2014).

Al-Ahmadi and Al-Ahmadi (2013) have studied the influence of altitude on the spring rainfalls using two statistical methods called: local geographically weighted regression (GWR) and global ordinary least square (OLS). GWR is a statistical method used for local spatial modelling considering the variation of the variables through the space. While OLS is a global regression assuming that the relationships between variables are constant over the whole region. The results of their study show that the application of GWR provides a better explanation of the relationships between the annual rainfall and the altitude. In addition, Tisu (2012) has explored three methods for yield estimation in Slovenia. Specifically, GWR, OLS and self-organizing map SOM the results show that GWR provides better result than OLS and the SOM poorly identified clusters in the data in his study.

Data mining techniques have proven their efficiency and flexibility in data analysis compared to statistical models (Hochachka et al., 2007). Particularly, data mining clustering technique called Self-Organizing Maps (SOMs) is used in a variety of disciplines. For instance, SOM have been used to identify a pattern of spatio-temporal disease dataset (Augustijn and Zurita-Milla, 2013). In addition, SOMs can be used for correlation hunting between variables (Vesanto and Ahola, 1999).In this research SOM used to cluster the spatial units that have similarity in the temporal behavior of the tick bites occurrence. Silva and Marques (2010), described the advantages of SOM which make it suitable tool for detecting correlation between the variables: On the one hand, SOMs have visualization capabilities make the interpretation of the result easier and richer. On the other hand, SOMs are considered a robust method to deal with the noise in the data (i.e. outliers).

2.6. The Self - Organizing Map (SOM)

The Self - Organizing Map (SOM) is a popular neural network algorithm, invented by professor Teuvo Kohonen in 1981-82 as a new visualization tool (Teuvo Kohonen, 2013). SOM has several properties which make it a powerful tool for clustering, recognizing pattern, visualizing, and analysing multi-dimensional datasets (Yin, 2008). On the one hand, SOM reduces the dimensionality of the data by representing a multi-dimensional data to a lower dimensional space, commonly a one or two dimensional space. On the other hand, it displays the similarity in the data by clustering. Therefore, it has been used for solving clustering problems and information extraction from complex datasets in variety of disciplines, for instance industry, natural science, finance, and engineering (Kohonen, 2013).

Typically, the SOM's structure consists of two layers: the input and output layers. Figure (5) illustrates the structure of the SOM as described in (Zhou et al., 2014) . Here, the input layer which is the one classified the input data based on their similarity; each black circle in the input layer represents an explanatory variable from the dataset. While the output layer (Kohonen map) of the SOM is consist of D number of neurons arranged in a 2- dimensional regular space. In between the adjustable weights (or network parameters) connect the neurons of the output layer with the every variables of the input layer. According to Tisu, (2012) "each neuron's properties are derived from all variables of input data during the process of training".

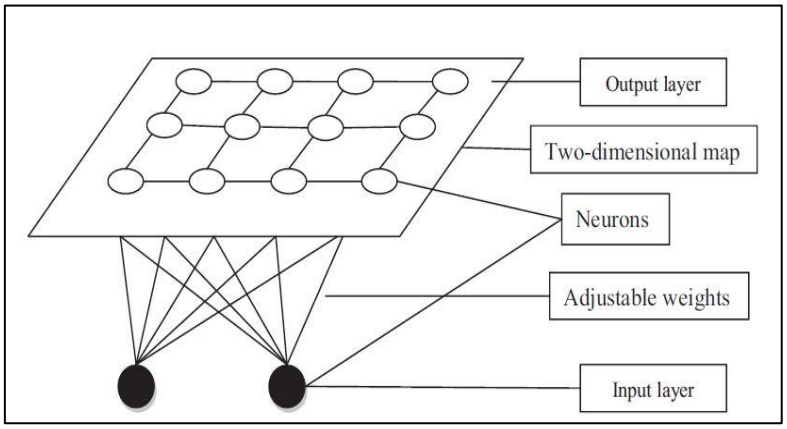


Figure 5: The typical structure of the SOM. Source: Zhou et al. 2014, p. 1474

SOM is unsupervised learning algorithm, learn without any previous knowledge. Therefore, no assumptions about the data are required (Ritter et al.,1992; Kaski et al.,1998; Chon, 2011). Kohonen, (2001) described the SOM as a nonlinear, order, smooth mapping of multidimensional input data onto units or regular low dimensional grid. Typically, the input dataset is represented by a matrix, consists of n rows and p columns. As an example: Table (1) illustrates the structure of an ecological data used by Giraudel and Lek (2001) for training a SOM in order to study the structure of ecological community. In this matrix the species types are represented by the rows while the sample units are represented by columns.

Table (1): SOM input data. Source: Giraudel & Lek (2001), p.330.

		Sample units			
		SU_1	SU_2	.....	SU_p
Species	Sp_1	X_{11}	X_{11}	.....	X_{11}
	Sp_2	X_{12}	X_{12}	.....	X_{12}
	Sp_3	...	...	.....	...
	Sp_4	Xn_1	Xn_2	.....	Xn_p

The SOM algorithm aims to project the input layer (which is the sample units in this example) onto a regular grid of neurons as an output layer. The implementation of the training process has been summarized in Vesanto and Ahola (1999) as following:

Initially, n-dimensional prototype vectors (weight vector) are assigned to each neuron in the SOM as: mi = [mi1 mi2 ... min] where is n represents the dimension of the input space. Iteratively, a random data sample x is chosen from the input dataset. By computing the Euclidean distance between the data sample x and the other weight vectors mc the best closest unit is found. This unit is called the best matching unit BMU. After that the weight vector of the BMU and its neighbour will be updated according to the formula:

$$mi = mi + \alpha(t)\, hci\,(t)(x - mi) \qquad (1)$$

Where α(t) is the learning rate and hci (t) is neighbourhood function and smoothing kernel centred in the best matching unit c. In this formula the value of (t) in neighbourhood function of is an integer which represents one step in the sequence (Kohonen, 2013).

The initialization and parameters of the SOM has an effect on the result of SOM. Therefore, analyst can try several initializations until achieve proper result, Commonly the quality of a SOM can be estimated by quantization and topographic error (Laaksonen and Honkela, 2011). The parameters are including: grid shape (map lattice), map dimension number of iterations learning rate and the number of neurons. Actually there are no strict rules to use for selecting the properties of SOM. Therefore, after checking the quality of the output after the first setting the analyst can decide by trial-and-error method which exact values must be selected. Nevertheless, there are several guidelines mentioned in literature can be used to select appropriate SOM configurations.

As mentioned before, neurons of the output layer are located on a regular space. The lattice of the grid has a hexagonal or rectangular shape as shown in figure (6). Usually, the map which is produced by hexagonal lattice is smoother and more pleasing to the eye; for this reason, the use of hexagonal grid is recommended (Kohonen, 2013).

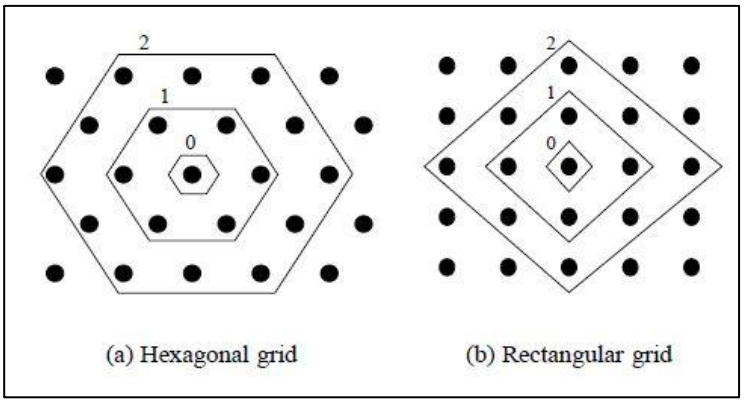


Figure 6: Different shape of the grid a) hexagonal lattice b) Rectangular lattice. Source: Vesanto, (2002).

Another important property of SOM is the map size (the dimension of the map). The map size should be suitable to identify the deviation of the input dataset; selecting too large map size may lead to over fit the models (Tisu, 2012). Moreover, Kohonen (2013) recommended that the relation of the map dimensions should be at the least follow the vertical and horizontal dimensions of the largest principal components of the input dataset. In general the oblong map is more efficient and faster in learning than the square one. Regarding to the selection of number of neurons in SOM there are no strict rules to follow, it may vary from few dozen to hundreds of neurons. To select the appropriate number of neurons, Park et al. (2014) used s heuristic equation (5* sqrt (n)) where n is the number of samples used for training. Finally, the number of iterations: according to Kohonen (1990), SOM learning is considered as stochastic process where the number of iterations has an influence on the mapping accuracy. To achieve sufficient statistical accuracy a large number of iterations are required. Commonly, at least 500 multiply by number of neurons in the map used to determine the number of iterations.

There are several methods for SOM's visualization. In this context, we will discuss two of them, which are Unified distance matrix and the component planes. In particular, unified matrix (U-matrix) is considered as one of the most used method in identifying clustering structure in SOM result (Vesanto and Sulkava,

2002). In its core, U-matrix visualizes the Euclidean distance between each neuron and its neighbouring neurons and shows the result in different colours as shown in figure (7). The clusters can be identified either automatically by the software (i.e. in Kohonen package) or by visual inspection, the analysts can easily recognize the clusters, but the result may differ from analyst to another (Tisu, 2012).

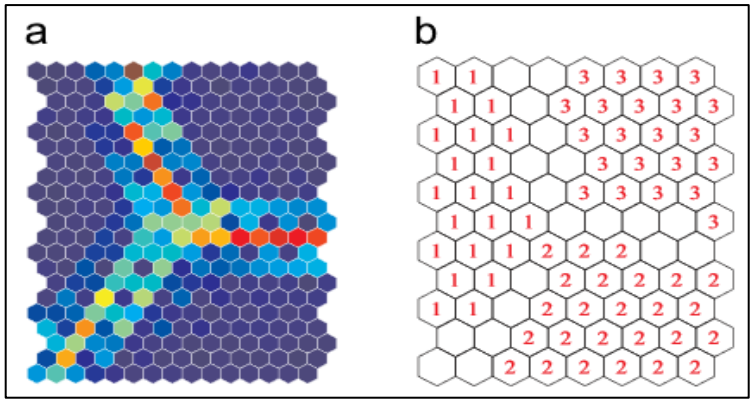


Figure 7: a) U-matrix and b) labels obtained by SOM for Iris dataset. Source: Kamimura (2010), p.2650

Typically, the structure of SOM is sliced into several component planes. Each component plane represents one variable in the dataset, and displays how the variable's values vary among the neurons as shown in Figure (8). The general variations of the variables patterns can be recognized in each part of the SOM. Thus, if two component planes show similar patterns, this means their corresponding variables are highly correlated. Moreover, the visual inspection to compare the pattern in the correlated component planes can lead to determine if they are positively or negatively correlated. However, this is feasible only in case of small dataset (Ejarque-Gonzalez and Butturini, 2014).

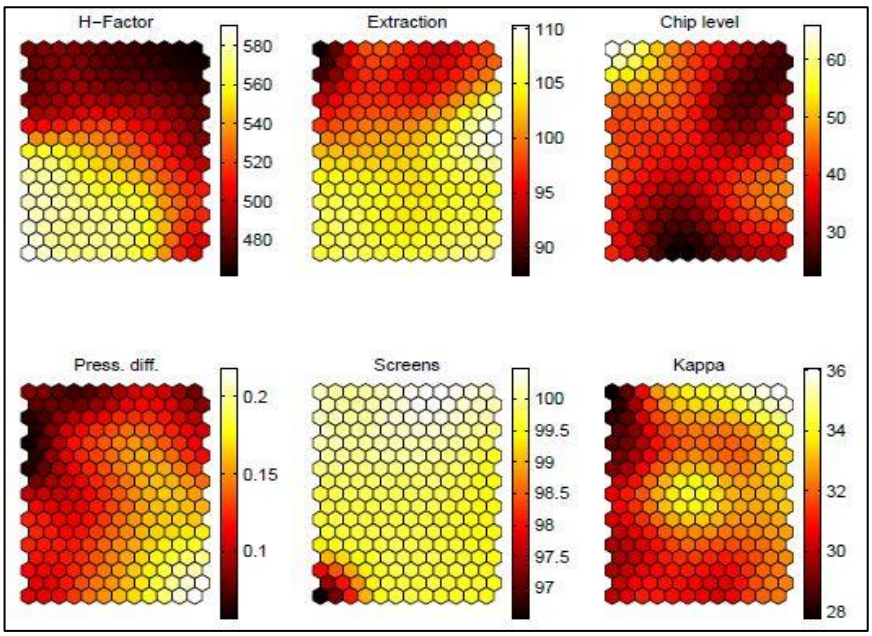


Figure 8: Component planes of SOM trained using digester data. Source:(Himberg et al.(2001), p.14

SOM has proven its efficiency in data mining and knowledge discovery in databases. Because it has the capability to recognize patterns, clustering and reducing the dimensionality of the massive datasets (Ejarque-Gonzalez and Butturini, 2014). In this research, SOMs was used for clustering of the spatial units which have similarity based on their temporal behavior of tick bites occurrence.

There are different packages to apply SOM such as: "SOM Tool box" in Matlab software, "class", "som", and "Kohonen" are available in R free statistical software (Tisu, 2012). In this research "Kohonen package" was used.

.

3. MATERIAL AND METHODS

This chapter provides a description of the datasets used in the analysis. In addition, it explains the methods selected to implement the proposed workflow phases and finally it answers the research questions. Thus, achieves the main objective of the research.

3.1. Data

Two types of datasets were used in this research. The first data type is a volunteered geographic dataset (tick bites dataset) in the Netherlands. The second data type is composed by a set of environmental variables relevant to this case study. The following subsections provide a description of these datasets:

3.1.1. VGI dataset

The "Tekenradar" and "Natuurkalender" are two Dutch projects linked to the phenological monitoring networks based on volunteers' contributions. In particular "Tekenradar" was launched in 2012 by Wageningen University and the National Institute for Public health and the Environment (RIVM). The project aimed to determine how often the occurrence of tick bites may lead to Lyme disease. To ease the achievement of this aim, Tekenradar allows for people to contribute in research about tick bites by reporting occurrence of tick bites(Large-scale study of preventive antibiotic usage against Lyme disease - Wageningen UR, 2013).

The tick bites dataset is a spatio-temporal dataset, consists of points representing tick bites incidences were reported by volunteers in different locations in the Netherlands over eight years, starting from 2006 until 2013. In the period 2006 up to 2013 the dataset was collected via the website "Natuurkalender", while the observations from 2012 and 2013 were collected via the website of "Tekenradar". Basically, the total number of records in the dataset consists of 27,947. However, the VGI datasets suffer from quality problems. Such as attributes incompleteness of the observations, positional error (e.g. tick bites located in water), and temporal accuracy. In this research all obvious mistaken observations were removed. Therefore, the total number of selected records used in the analysis was 22,778. The tick bites dataset has several attributes as shown in table (2):

Table (2): Tick bites dataset attributes.

Id	Attribute	Id	Attribute
1	Unique ID	7	The municipality sub-division
2	The absolute location (X,Y)	8	The Street address
3	Tick bite's occurrence date	9	The Postal code
4	Tick bite reporting date	10	The environment type (such as forest, garden…etc.)
5	The province	11	The activity (such as walking, gardening)
6	The municipality	12	The Volunteer's year of birth.

3.1.2. Environmental datasets:

Based on the literature review the key environmental variables for the case study were defined. According, the theoretical understanding of the phenomenon and the availability of the data, the following variables in Table (3) were selected to be used in the analysis:

Table (3): Selected environmental variables

Variable	Source	Reasoning
Land use Land cover (LULC)	PDOK	Land use land cover can be related to tick availability and human activity
Forest mean patch area	Derived from LULC data	The mean patch area used to approximately assess the forest fragmentation phenomenon. Several studies linked forest fragmentation to the tick abundance.
Temperature	KNMI	Temperature has impact on ticks activity and development. And it could give an indicator of human recreation behavior.
School vacations	Web page of Dutch Ministry of Education, Culture and Science. And another relevant webpage	More recreation activities are expected during the schools' vacations periods.

The study area was divided to regular equal size spatial units (5km×5km) and all of these variables were extracted for each spatial unit. The Following subsections provide description of each dataset

Land use land cover dataset:

The Netherlands Land use land cover map (BBG 2008) was downloaded from The Dutch National SDI **(**PDOK**)** and reclassified in to eight separate layers in order to study whether the occurrence of the phenomenon varies among different LULC type:

- *Built-up*: this layer includes all built up areas, buildings, boarder land, green houses, and airport.
- *Agriculture*: all agriculture areas in The Netherlands were reclassified as one class.
- *Forest*: this layer includes all the forests.
- *Transportation:* This layer includes the roads network and railways.
- *Dry terrain:* includes all natural features which are classified as dry natural terrain.
- *Wet terrain:* includes all natural features which are classified as wet natural terrain.
- *Water:* includes all water bodies.
- *Recreation:* includes all recreation sites in the Netherlands

Further, the percentage of each LULC type in each spatial unit was calculated using a code written in Python programming language, see Appendix (C).

Forest mean patch area:

Forest mean patch area was derived from the LULC dataset. The forest layer was extracted and corrected using multipart to single part tool in ArcGIS10.2 to be sure about the number of patches in each spatial unit. Then the mean parch area was computed. Here the value of mean patch area was used to approximately evaluate the forest fragmentation phenomenon in each spatial unit.

Schools' vacations dataset:

The schools' vacations information and dates were collected from the official webpage of Dutch Ministry of Education, Culture and Science and other relevant websites[1]. In the Netherlands schools have four types of vacations: spring, May, summer and autumn vacation. The dates of these vacations are slightly vary from region to region all over the country Figure (9). Essentially, the schools are divided in to the following three regions:

Figure 9: Schools regions in the Netherlands.

- **North region**: this region includes all schools belonging to the municipalities in North-Holland, Friesland, Drenthe Groningen Overijssel Flevoland (except Zeewolde), Utrecht (Eemnes) and Gelderland (Hattem).
- **Middle region**: this region includes all schools belonging to the municipalities in South Holland, and the North Brabant municipalities of Werkendam and Woudrichem
- **South region**: this region includes all schools belonging to the municipalities in Limburg, Zeeland and North Brabant (except Werkendam and Woudrichem)[2].

[1] http://mens-en-gezondheid.infonu.nl/kinderen/3327-schoolvakanties-in-nederland-tm-2011.html
[2] http://southholland.angloinfo.com/information/family/schooling-education/school-holidays/

The dates of schools vacations were converted to a week of the year using timedate module in python programming language.
Table (4) shows sample of the data after this conversion:

Region	Autumn	Spring	May	Summer
North	43	8	18	Primary : 28-33 Secondary : 28-34
Middle	42	8	18	Primary : 30-35 Secondary : 29-35
South	42	9	18	Primary : 27-32 Secondary : 27-32

Table (4): Sample of school's vacations dataset for year (2012/2013).

Meteorological data

The Royal Netherlands Meteorological Institute (KNMI) is the national institute for weather forecasting. The principal tasks of KNMI include daily weather forecasting, monitoring the climate changes and serving seismic activity. In addition, this institute provides the public and government organizations and public community with daily weather data. For instance, daily temperature, precipitation and humidity (KNMI, 2015).

In this research, KNMI is our source for the maximum daily temperature. A time series of temperature for seven years (2006-2012) were available in raster format. The temperature data was processed by applying several steps using code written in python. First, the daily temperatures were aggregated to a mean weekly temperature. Second, a regular grid (5km×5km) was used to extract the mean temperature for each cell (spatial units). Finally, the data combined together in a suitable structure in space and time.

Environmental datasets preparation

Several steps were applied to prepare the environmental datasets. Data preparation required to structure the data in space and time to be ready for the correlation analysis. An overview of the preparation steps are summarized in the following Table (5)

Table (5): Environmental data preparation steps in the research.

Dataset	Preparation	Output
Land use land cover BBG2008	- Reclassified in to eight classes (mentioned in section(3.1.2) - Intersect with a grid (5km×5 km) - Calculate the percentage of each land use land cover type in each spatial unit	Grid shape file Cell size: 5km×5 km Attributes: LULC percentage in each cell(spatial unit)
Forest mean patch area	- Extract forests layer from LULC dataset - Correct Forest layer (using multipart to single part tool in ArchGIS10.2) - Intersect with a grid (5km×5 km) - Calculate Mean patch area	Grid shape file Cell size: 5km×5 km Attributes: mean patch area
School's vacation	- The raw data of the school vacations collected and converted to the week of the year instead of date. - This data formulated as a matrix, the rows represent spatial units, and the columns are week of the year. In this matrix value 0 was assigned for the weeks have no school vacation, value 1 was assigned for the weeks when primary or secondary school has vacation, and value 2 was assigned when primary and secondary schools both have vacation. - The matrix was Joined with schools regions map - Intersect with a grid (5km×5 km)	Grid shape file Cell size: 5km×5 km Attributes: school Vacation time as week of the year
Daily maximum temperature	- Aggregated weekly - Intersect with a grid (5km×5 km) - Average of temperature in each spatial unit calculated.	Grid shape files (5X5 km cell size). Attributes: weekly maximum temperature

3.2. Methods

This section describes the implemented methods to achieve the main objective of this research. Initially, several relevant literatures were reviewed to identify the more appropriate methods to be applied. In addition, to determine the key environmental variables those are relevant to this case study.

This research was carried out by applying the workflow illustrated in Figure (10). The workflow consists of three main stages:

- Data preparation and aggregation of the VGI dataset and the other environmental datasets. During this stage, spatio-temporal explorations were applied; we tested several aggregation levels to select proper analytical scale for the analysis. This stage was followed by data transformation stage.
- Data transformation stage, specifically this stage was applied for the tick bites dataset. Here, the temporal behavior of the ticks' bites occurrence was constructed as a new characteristic of the dataset. The temporal behavior was constructed to avoid the significant differences in number of observations from year to year. Moreover, it was used as similarity measure in the data mining stage.
- Data mining (clustering) and correlation analysis were implemented. Particularly, self-organizing map was selected to cluster the spatial units according to the temporal behavior of tick bites occurrence. Then we addressed the correlation between the occurrence of tick bites and the environmental variables. Finally the research work end by result analysis and discussion. The analysis was implemented using Python programming language and R software; programming was required to facilitate testing several parameters and aggregation levels. The following sections discuss each stage in the workflow in details.

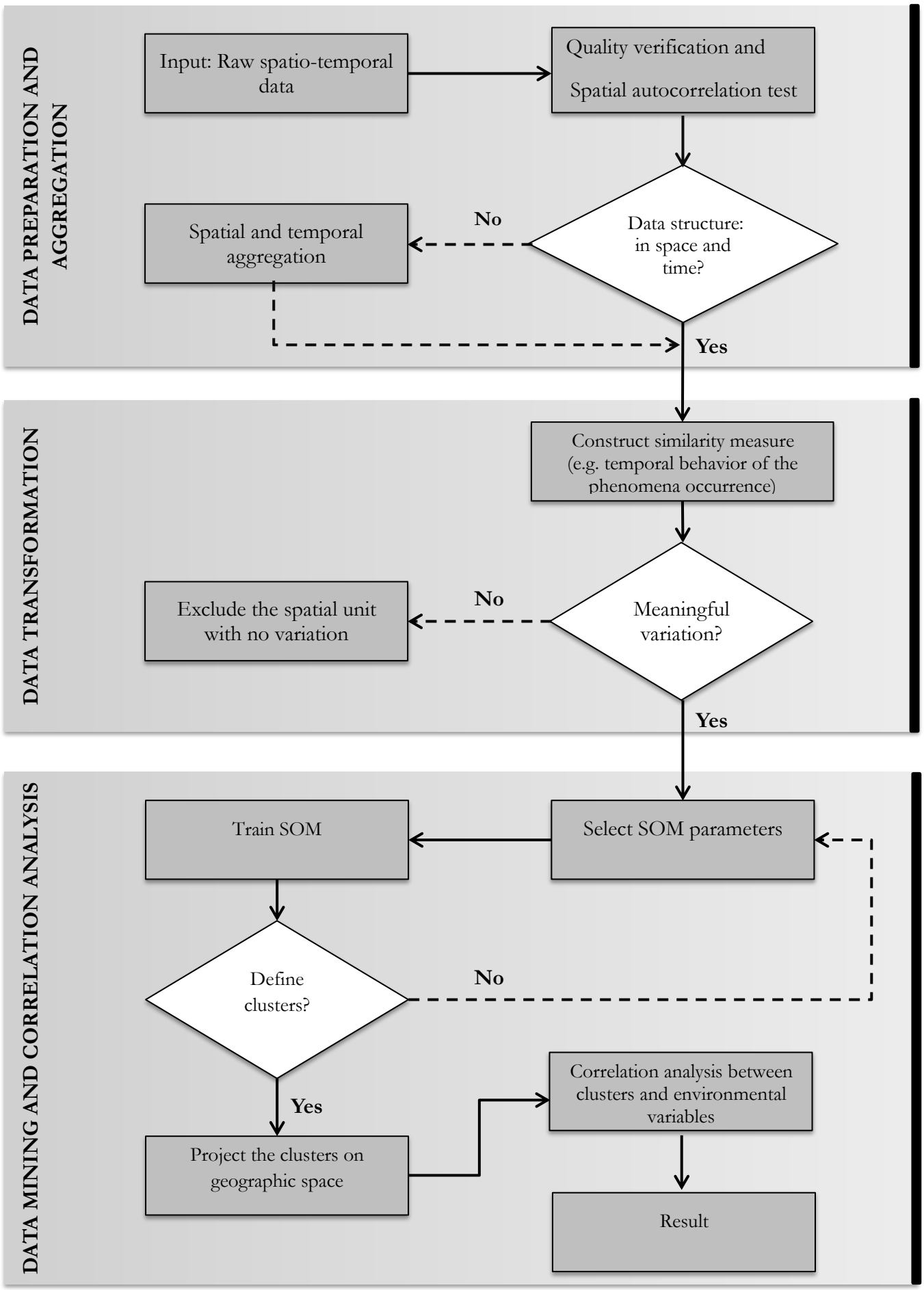


Figure 10: The Developed workflow for research implementation.

3.2.1. Data preparation and aggregation

Data preparation and aggregation stage was the most important stage in this project, aimed at making the required dataset fit for use in our analytic workflow. Several steps were applied to pre-process the datasets and make it in a proper structure for knowledge discovery process. The preparation of the environmental datasets was discussed in section (3.1.2). Here, we addressed the preparation stages of the VGI dataset (tick bites dataset). As shown in Figure (11), several preparation stages were applied to verify the quality of the raw VGI data, in addition to selection of proper analytical scale. Then, the data was aggregated in space and time to make it ready for the next stages which are data transformation and data mining. The following sub sections provide description of each stage in the workflow.

Figure 11: The over view of data preparation and aggregation stage.

Quality verification of VGI data (tick bites dataset)

As mentioned in section (3.1.1), there is always concern regarding the quality of the data collected by volunteers. Therefore, the quality of the tick bites dataset was verified and only the desired observations that have required attributes and located in the Netherlands were selected for the analysis. The quality verification was performed in ArcGIS 10.2, for this purpose the raw tick bites dataset was mapped into geographic space and explored in ArcMap. Specifically, the quality check considered three aspects, which are: data completeness, spatial location accuracy and temporal outliers. Therefore, all of the following obvious erroneous observations were removed:

- All observations which have no absolute location were removed.
- All observations located in water were removed.
- All observations located outside The Netherlands were removed
- All observations that out of the study time span (2006-2013) were removed.

After the quality check, the cleaned dataset was split into yearly files to be used in the second step of data preparation. In fact, tick bites dataset has point observations distributed through the space. So we need to determine how these individual observations are distributed in the space. In short, are they clustered or randomly distributed in space? This is required to select proper analysis method. To achieve this aim the following test was executed:

Spatial autocorrelation test

Spatial autocorrelation is a common phenomenon exists in ecological and environmental spatial data where the objects have specific locations. Specifically, spatial autocorrelation defined as similarity or correlation between values (Stojanova et al., 2013). In general, the nearby objects in the dataset are more similar to each other than those objects which are a part in space. Actually, this can be justified by the first law of geography Tobler's first law, which states: "*everything is related to everything else, but near things are more related than distant things.*" The spatial autocorrelation indicates to which degree nearby objects are similar to

each other. Normally, positive spatial autocorrelation exists in the dataset if the features are clustered to each other. In contrast, negative spatial autocorrelation exists when the features are dispersed in space. And the spatial auto correlation absent if the objects are independent of each other.
Statistically, the existence or absence of the spatial correlations among the observations can be examined by several methods such as: Geary's C, Getis's G and Moran's I test. In this research Moran's I was selected to test the existence of the spatial autocorrelation among the observations of the Ticks dataset. Moran's I test was performed in ArcMap for Ticks dataset yearly. Actually, it is important to test if the spatial autocorrelation exists within the data or absent. This can help the analyst to choose the proper analytic method to deal with the problem. Commonly, statistical methods assume that the observations are independent. The existence of spatial autocorrelation violates this assumption(Dale & Fortin, 2009). And this supports the choice of applying data mining instead of statistical methods to deal with this research problem.

Data aggregation

Data aggregation groupings individual observations in the dataset into subsets, this reduces the amount of data to be analysed (Andrienko and Andrienko, 2006) . The data aggregations based on the analytic scales are selected by the analyst. The choice of spatial and temporal scale is a very important task. Because, the discovered patterns in the datasets are directly influenced by the spatial and temporal scale. In fact, the finer the resolution the more likely to discover interesting patterns in the dataset (G. Andrienko et al., 2006) . But it is not always the case.

This stage of the workflow was intended to structure the dataset into space and time matrix (S×T). To achieve this aim, coding was used to implement the spatial and temporal aggregation of tick bites dataset. Initially, the study area was divided to equal sizes sub regions using a regular grid. Then, the numbers of observations in each spatial unit were computed. For better understanding of the phenomenon and select an appropriate spatial analytical scale, we tried variety of aggregation levels. As the aggregation code allows for the analyst to change the aggregation level easily. The aggregation level that clearly represents the data was selected.

Each observation in the tick bites dataset has a "bite occurrence date" as an attribute. This, attribute was transformed by converting its date to a week number (WOY). This makes the data more suitable for knowledge discovery process. This conversion was implemented using datetime module in Python (see Appendix C). Further, the numbers of ticks' bites in each week were computed. This choice can be justified by the knowledge that a clear variation can be found from week to another. In addition, this aggregation reduced the number of temporal units to be 52 per year instead of 365 per year (in case of individual observations). At the end of this stage tick bites dataset became structured in space and time (S×T) as shown in Figure (12).

	Temporal units					
Spatial unit ID		w1	w2	w3	…….	w52
	1					
	2					
	3					
	…					

Figure 12: Data structure in space and time matrix.

Spatio- temporal exploration

Data exploration is a preliminary step in this research. The tick bites data set was explored spatially and temporally using maps and exploratory data analysis (EDA). This stage of the analysis attempts to detect three issues:

- Temporal variation of number of reported tick bites from year to year
- Temporal variation of number of reported tick bites during every year.
- The change of the spatial distribution of the reported ticks bites in The Netherlands.

Here, we addressed the third question of this research about defining the high-risk period during the year. To achieve the answer of this question tick bites dataset were explored using histograms as an exploratory analysis tool to illustrate the temporal variation of tick bites occurrence and highlight the high risk period during the year.

3.2.2. Data transformation

Data transformation stage was intended to construct a similarity measure for comprehensive data analysis. The spatial units will be clustered according to this similarity measure. Here, the temporal behavior of the tick bites occurrence was selected as the similarity measure for the clustering. This step was implemented as follows: the cumulative sum of the observations (reported ticks bites) was computed. Further, we found the week of the year when (25%, 50%, 75%, and 100 %) percentages of the observation were reached. A simplified example explains how the code created the similarity measure in this research. The code is provided in Appendix (D):

Simplified example:

Step (1): The first two columns in Table (6) represent the original data related to an arbitrary selected spatial unit (x). Based on this data the values of the cumulative sum of the observations were computed weekly.

Table (6): Sample data of an arbitrary selected spatial unit.

Week	No. of observations	Cumulative sum (CumSum)
1	0	0
2	0	0
3 . .	1	1
20	5	6
21	6	12
22 . .	3	15
33	7	22
34	3	25
. .		
52	0	30

Step (2): To represent the temporal behavior in each spatial unit, by coding we found out in which week (25%, 50%, 75%, and 100%) of the cumulative sum of the observations were reached. Then the minimum absolute differences between these values and the cumulative sum were computed. Finally the code read the index of that values which represent the week of the year (WOY) when these percentages were reached. Table (7) explains the calculations of the temporal behavior of tick bites occurrence in the spatial unit (x)

$$A = 0.25 \times CumSum[52] = 7.5$$
$$B = 0.50 \times CumSum[52] = 15$$
$$C = 0.75 \times CumSum[52] = 22.5$$
$$D = 1.00 \times CumSum[52] = 30$$

Table (7): Calculations of the temporal behavior.

Week	No. of observations	Cumulative sum (*CumSum*)	$\lvert A - CumSum \rvert$	$\lvert B - CumSum \rvert$	$\lvert C - CumSum \rvert$	$\lvert D - CumSum \rvert$
1	0	0	7.5	15	22.5	30
2	0	0	7.5	15	22.5	30
3	1	1	6.5	14	21.5	29
.						
.						
20	5	6	**1.5** ←	9	16.5	24
21	6	12	4.5	3	10.5	18
22	3	15	7.5	**0** ←	7.5	15
.						
.						
33	7	22	14.5	7	**0.5** ←	8
34	3	25	17.5	10	22.5	5
.						
.						
52	0	30	22.5	15	5	**0** ←

As a result, 25%, 50%, 75%, 100% of the observations in this spatial unit occurred in the weeks 20, 22, 33, 52 respectively as shown in the following Figure (13). These calculations applied for all spatial units.

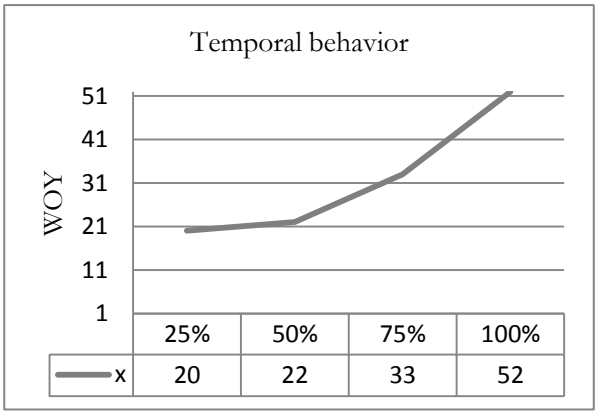


Figure 13: Temporal behavior of tick bites occurrence in an arbitrary selected spatial unit (x).

3.2.3. Data mining

After the data transformation stage the dataset is ready for data mining. Self-organizing maps (SOM) were selected as a data mining tool. Previously, SOMs were used as classification technique applied on various phenomenon (Tisu, 2012). In this research, SOMs were applied to cluster the spatial units which are similar in the temporal behavior of the Ticks bites occurrence. The reason behind this choice can be summarized by:

- SOM is able to cluster the input data according to their similarity as stated in the literature review (section2.6). Here, SOMs were applied to discover whether the temporal behaviour observed in the spatial units define clear clusters. In addition, the spatial units within each cluster are supposed to have similarity in the environmental variables. Therefore, studying the explanatory variables within each cluster can highlight which environmental variable that has an influence on the temporal behavior of the phenomenon.

SOM training

In this analysis, SOM was trained based on the complete tick bites dataset. The SOM's parameters were selected based on the knowledge gained from the literature. Here, we briefly describe our choice for the SOM parameters:

Map dimension: the selected map size and shape should be suitable to avoid models over fitting problem. According to the literature, the oblong map is faster in learning and more efficient than the square map. In addition, Park et al., (2014) determined the appropriate number of neurons by ($5*\sqrt{n}$) where n is the number of samples. In this research we tried several map sizes and the size (10×15) was selected, because it shows results better defined clusters.

Neighbourhood shape: the lattice shape could be rectangular or hexagonal, the hexagonal lattice is highly recommended in literature, and for this reason hexagonal lattice was selected in this analysis.

Number of iteration: for better mapping, a large number of iterations are required. As recommend in (Kohonen, 1990) the number of iterations should minimum 500 times the number of neurons. In this analysis the selected number of neurons is 150 neurons; according to this fact the number of iterations should be at least 75,000 iterations. For better mapping we set the number of iterations to 100,000 iterations.

Learning rate: according to Kohonen, (1990) the value of the learning rate is $0< \alpha(t) <1$, and its value always decreasing. In this research we set the value of learning rate starting 0.05 decreased to 0.01.

SOM clustering and its projection to the geographic space

After the completion of the SOM training using the previously mentioned parameters and based on the tick bites dataset, the output plots of the SOM such as U-matrix, distance plot and count plot were visually investigated and interpreted, to assess the result. Based on the visual inspection of the U-matrix which represents the distance between each neuron and its neighboring neurons, the number of clusters was selected. Precisely, we set the number of clusters to be four. In order to assign each neuron to a cluster an automatic clustering method called hierarchical clustering was applied. This process aimed to group the spatial units which have similar temporal behavior of the tick bites occurrence in one cluster. SOM training and clustering were implemented in R software using "Kohonen" package.

Based on the result of defining the clusters, the SOM output was separated yearly and projected into the geographic space in ArcGIS 10.2. Eight maps were displayed to assess the spatial pattern of the clusters and to evaluate the relation between the clusters and the environmental variables in different geographic location.

The relation between the clusters and the environmental variables

This section covers the study of the relationship between the SOM clusters and the environmental variables. This was implemented by doing statistical calculations for these variables within each cluster. According to these calculations we assessed whether these environmental variables have variations within the clusters.

Environmental datasets were prepared as mentioned in section (3.1.2). As a result of the preparation stage the mean temperatures, percentage of LULC, mean patch area, and school vacations weeks were defined in each spatial unit. All of these variables were calculated within each cluster to assess if there is clear variation within the clusters and if any of these variables can be used to predict the temporal behavior of the tick bites occurrence.

4. RESULT AND DISCUSSIONS

The previous chapter discussed the workflow used to discover possible correlations between the temporal behavior of tick bites occurrence and the selected environmental variables. Here, this chapter presents the results of implementing each stage in the workflow of this research. Starting, from the results of the data preparation and aggregation stage, followed by the results of the data transformations. Finally, results of the data mining and correlation analysis were discussed.

4.1. Data preparation and aggregation

The data preparation stage includes three steps: VGI data quality verification, data understanding and aggregation, spatial and temporal exploration. The following subsections will discuss the results of implementing all of these steps.

4.1.1. VGI quality verification

The original tick bites dataset contains 27,947 records. These records are representing tick bites incidences reported by volunteers in the Netherlands. In fact, the quality problems are common in the VGI datasets. Therefore, the data preparation stage was started by investigating the quality of tick bites dataset. The quality aspects those we considered are shown in Figure (14): the records that have incomplete attributes (specifically, location) were excluded from the dataset. In addition, all the obvious positional mistakes were removed, such as tick bites reported in water or outside the Netherlands. Moreover, tick bites dataset contains observations were reported out of the temporal span of this research (2006-2013), all of these observations were excluded.

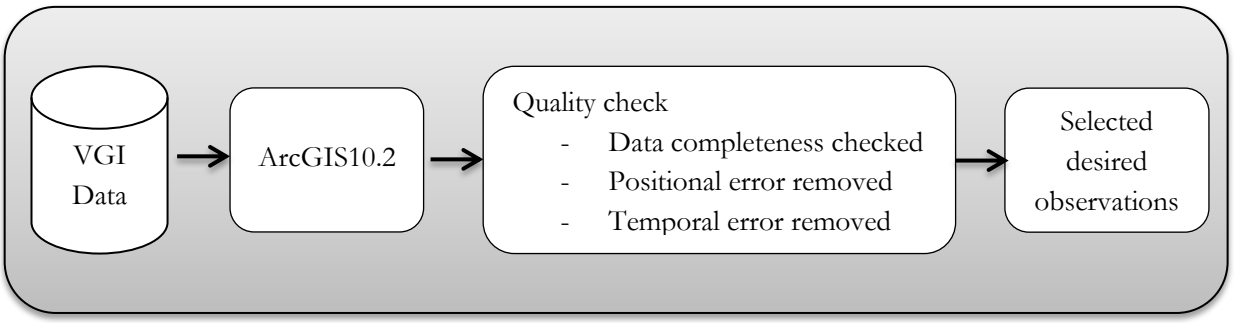


Figure 14 : Overview of the VGI data quality check.

The final number of the selected observations was 22,778 tick bites incidences were reported through eight years (2006-2013). Actually, the number of the reports varies from year to year and the last two years show significant increase of the numbers of reported tick bites as shown in Table (8).

Table (8):Number of reported tick bites (yearly).

Year	2006	2007	2008	2009	2010	2011	2012	2013
Number of tick bites	1858	1555	1007	2186	1084	1182	6292	7632

Previous studies shown a significant increase in the number of Lyme disease incidence last decade in The Netherlands and in Europe in general (Hofhuis et al., 2006, Tijsse-Klasen et al., 2010). And the tick bites dataset indicates an obvious increase in the number of tick bites occurrence in the Netherlands.

The significant increase in the number of reports through Tekenradar can be justified by the following: First, the public health organizations attempted to pass messages through media to increase the awareness about the risk of the tick bites and tick-borne diseases on public health. Second, this came with the wide spread of the smart devices which facilitated the user to create content via the Internet and this motivates the volunteers' contribution in reporting the occurrence of tick bites. Third, could be due to the change of the geographic distribution of ticks in the Netherlands over the recent years that increased the chance for the human to get a tick bites.

In a deeper exploration for the tick bites dataset it was found 13326 of the volunteers were reported their year of birth. The age of the person who has been bitten was computed and displayed using a histogram shown in Figure (15). From the figure, one can recognize two peaks, indicating that the majority of the volunteers were old people in age approximately 62.5 years old. Moreover, there are significant numbers of the people who have been bitten were children less than ten years old. This indicates that children have high risk of Lyme disease and this knowledge mentioned in (Beaujean et al., 2013). And this supports our choice to include the school vacations as an exploratory variable in the analysis.

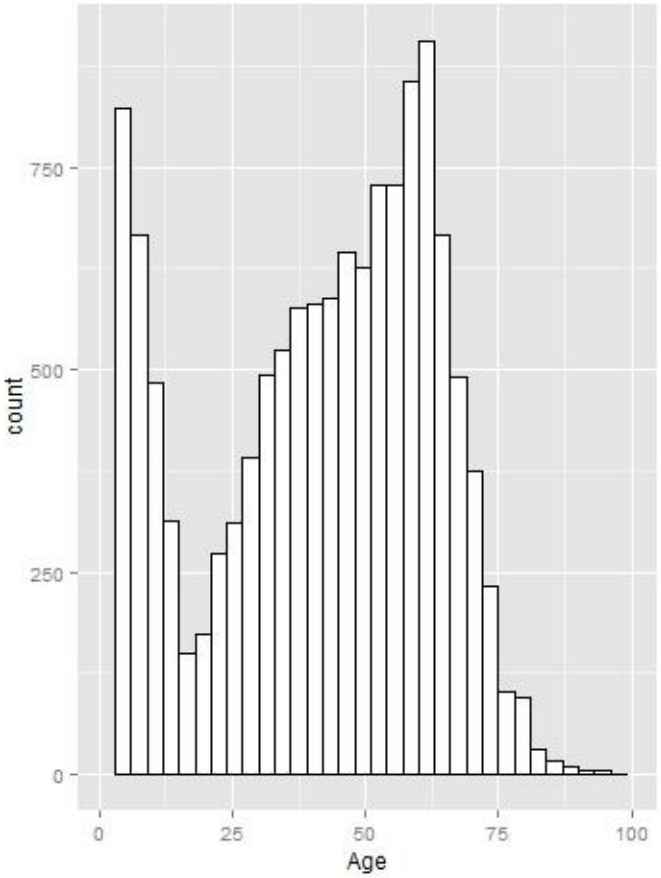


Figure 15: Number of the Volunteers according to their age.

4.1.2. Spatial autocorrelation test

This step of the workflow aimed at investigation and understanding of the spatial pattern of the phenomenon in the geographic space. To achieve this goal, Moran's I test was selected to test the spatial autocorrelation among the tick bite dataset observations. In order to perform Moran's I test, tick bite dataset was separated per year, and the test was executed for each year separately. Moran's I test was executed using *analyzing patterns* tool box in ArcGIS10.2

The result of Moran's I test was revealed on existence of positive spatial autocorrelation among the observations of tick bite dataset for all years except 2006 and 2011(see Appendix A). According to the P-values, the results showed that positive spatial autocorrelation is exist among the observations of years (2007, 2008, 2009, 2010, 2012, and 2013), this indicates that the observations in the related data are spatially clustered. In contrast, datasets of year 2006 and 2011 showed randomness in their distribution this could be because less observations. The importance of this test was to guide us to select proper analysis method. The existence of spatial autocorrelation among the observations violates the assumption of statistical analysis methods which assume the randomness and dependency of the observations across the geographic space. Therefore it became inconvenient choice for the analysis.

4.1.3. Data aggregation

The data aggregation stage was required to structure the tick bites dataset in space and time (S×T) matrix, two types of aggregation were implemented temporal and spatial aggregation. The following subsections describe how these steps were applied.

Temporal aggregation

The temporal aggregation step was intended to group the observations weekly, to reduce the number of the temporal units. Specifically, the week level was selected as a temporal unit in this analysis; because a clear variation in the number of observations and the environmental variables was expected. The temporal aggregation was implemented as follows: the week of the year (WOY) was generated for each observation based on one of its attribute "tick bite's occurrence date". These dates were converted a week number using datetime module in Python. Then, the number of observation computed weekly.

To investigate the temporal variation in the observation between the years, the tick bites dataset was displayed using different graphical representations, this can provides more details about the dataset. In Figure (16) the dataset was displayed using boxplot per year. From this figure, one can observe that the mean of WOY exists in between week 20 (mid of May) and week 30 (end of July) every year. Moreover, there was slight difference between the observations of each year. For instance, in the year 2007 reporting ticks bite was started relatively earlier compared to other years. According to Gray et al. (2009), they mentioned that the seasonal questing activity of ticks is sensitive to temperature, this guided us to investigate the weekly minimum temperature for all years. And we noticed that the average minimum temperatures in the first three weeks of year 2007 were (5.65, 8.20, 6.18 C^{o}) respectively. These temperatures are higher than usual; this implies that year 2007 had warmer winter than other years. The warm winter could lead ticks to start questing earlier, and this could be the reason behind the early reports of tick bites in year the 2007. The Figure(16) shows very early observations : on one hand, these could be erroneous reports, on the other hand, it could be true, as mentioned in the literature, *Ixodes ricinus* is the most common tick species in Europe (Reye et al., 2010). And according to Estrada-Peña et al. (2004) "*Ixodes ricinus* lives in a wide range of temperatures and environmental conditions throughout Europe" this can justify the occurrence of tick bites out of the tick questing season and in the early weeks of the year .

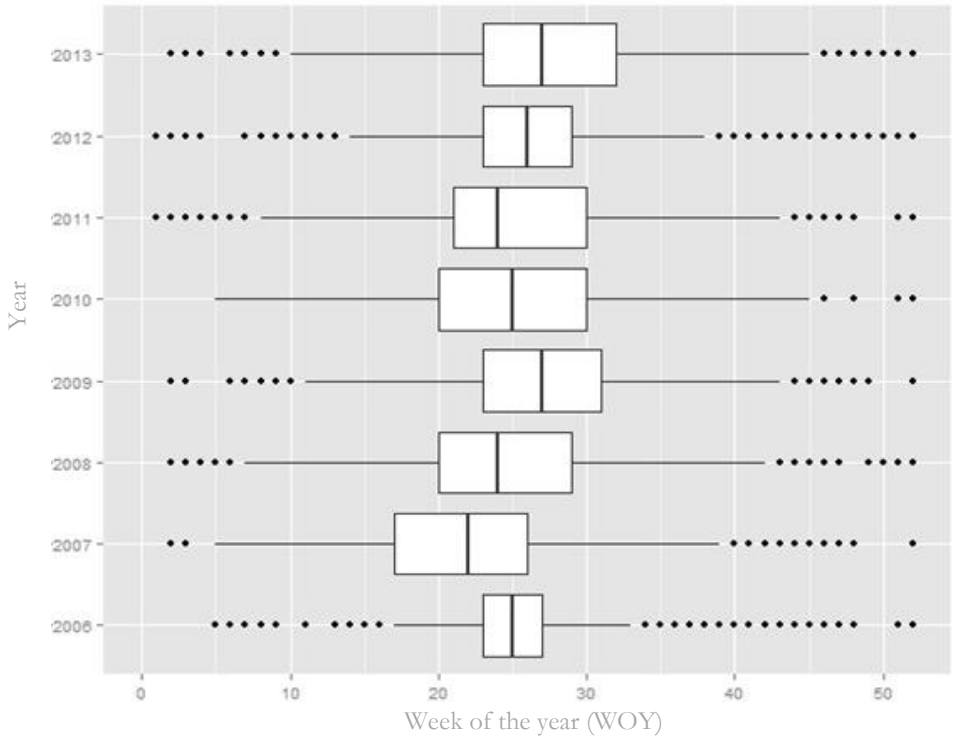


Figure 16: Boxplots for the reported tick bites per year.

In this stage we address the third research question (section1.2.2), regarding the high risk period during the year. For this purpose additional exploration was required, the frequency histogram was selected as an EDA tool to highlight the period that has a significant number of reported tick bites every year. Figure (17) and Figure (18) illustrate the weekly number of reported tick bites per year in the period (2006 – 2013). From the histograms, one can clearly recognize there is a variation of the number of observations from year to year. Specifically the last two years (2012-2013) have the highest number of observations (6292, 7632 reported tick bites) respectively. While, year 2008 has the lowest number of observations (1007 reported tick bites).

The peak of the histograms occurred in between week 20 and week 30 in all years except year 2007 and 2009. In 2007 the peak relatively occurred earlier than other years, based on the temperature dataset, year 2007 was warmer than other years. In contrast, year 2009 had colder winter than usual, and we notice the peak occurred relatively late few weeks after week 30. This indicates that a very cold winter which could be the reason that led to a delay of questing activity of the ticks in that year.

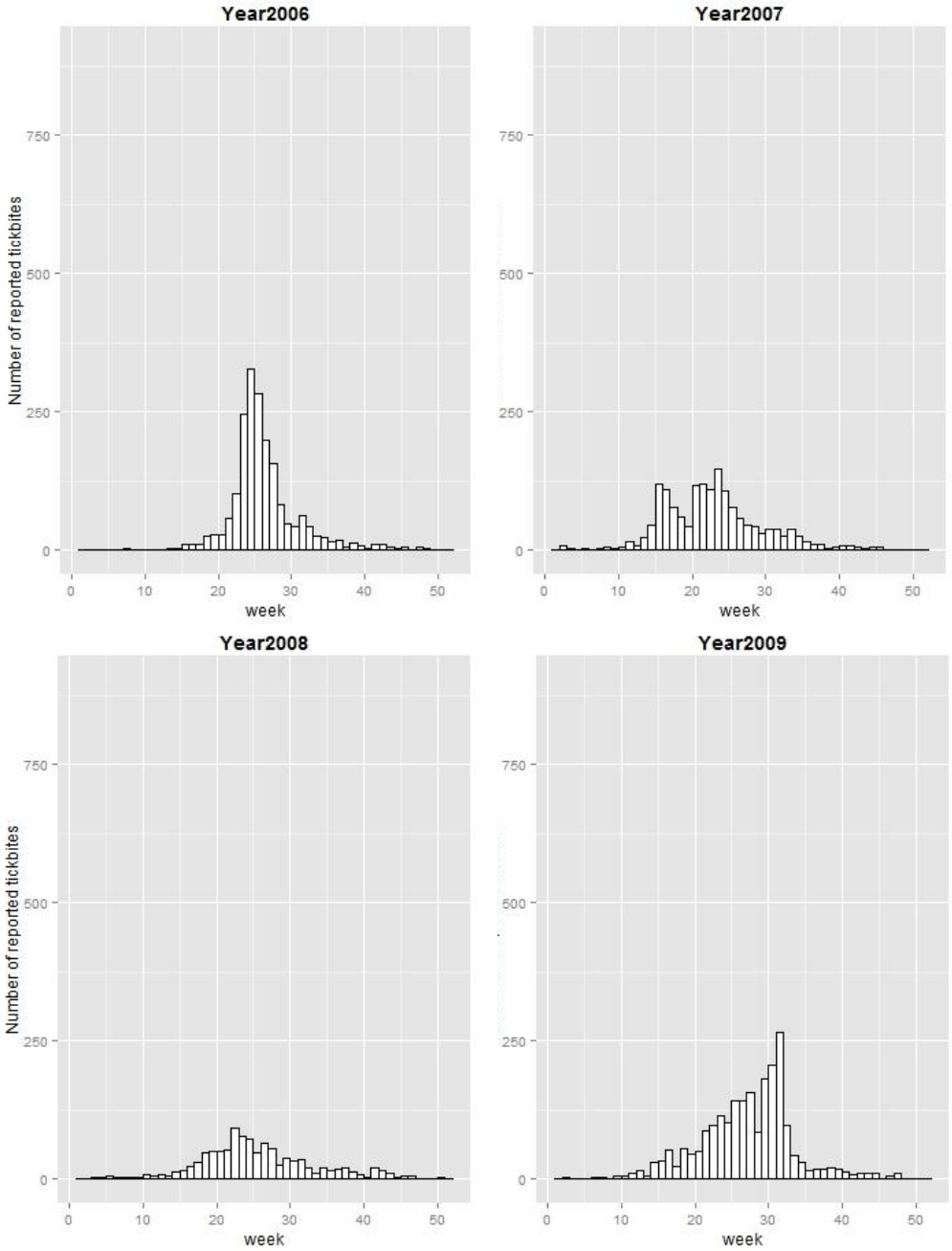


Figure 17: Histograms of the weekly number of reported tick bites per year(2006-2009).

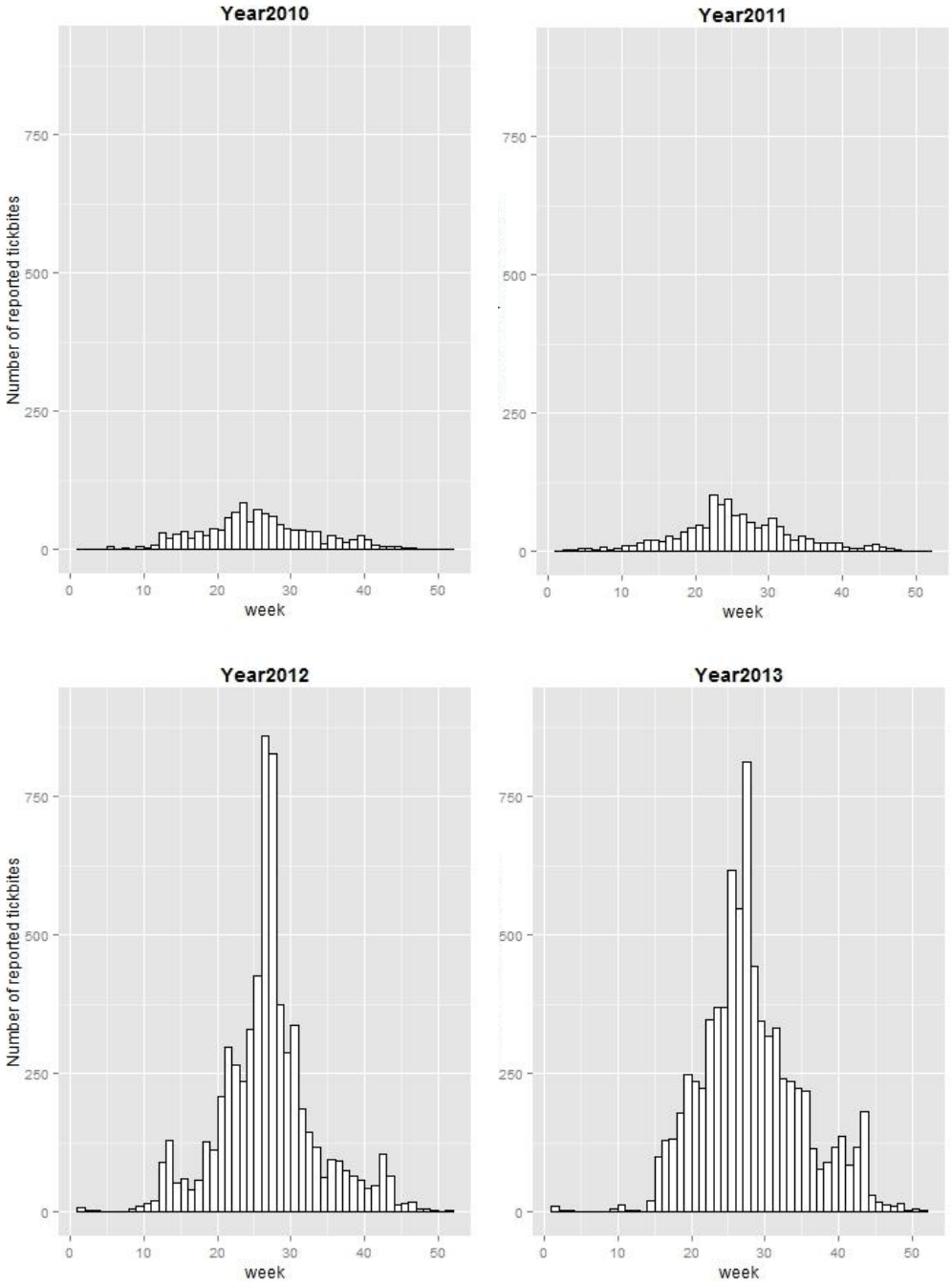


Figure 18: Histograms of the weekly number of reported tick bites per year (2010-2013)

Actually there is a clear increase in the number of reported tick bites can be detected from the histograms of year 2012 and 2013. Moreover an autumn peak is obvious in both histograms, which is usually smaller than the spring peak (May to July in this study) (Estrada-Peña et al., 2004). The reason for the smaller peak is due to only adult ticks are questing in autumn, which are less in number than nymphs and larvae which are not active during this time. Furthermore, autumn peak coincides with autumn school vacation in the Netherlands, therefore human recreation activities are expected during this period. This could increase human exposure to tick bites. And this motivated our choice to include school vacation time as an exploratory variable in this analysis.

Spatial aggregation:

This section discusses the aggregation process of point observation to an area. Raster analysis was applied in this research. The study area was divided in to an equal sizes regular grid, and the numbers of point observations were counted in each grid cell. For implementing this step, a spatial aggregation code was prepared in Python. This code allow for the analyst to try several aggregation levels easily. Here in this step the second research question concerning the proper aggregation level was addressed here. In fact, the selection of the aggregation level affects the discovered pattern and the result of the analysis. Therefore, several aggregations levels were investigated to decide which aggregation level is appropriate for implementing this analysis as shown in figure (19).

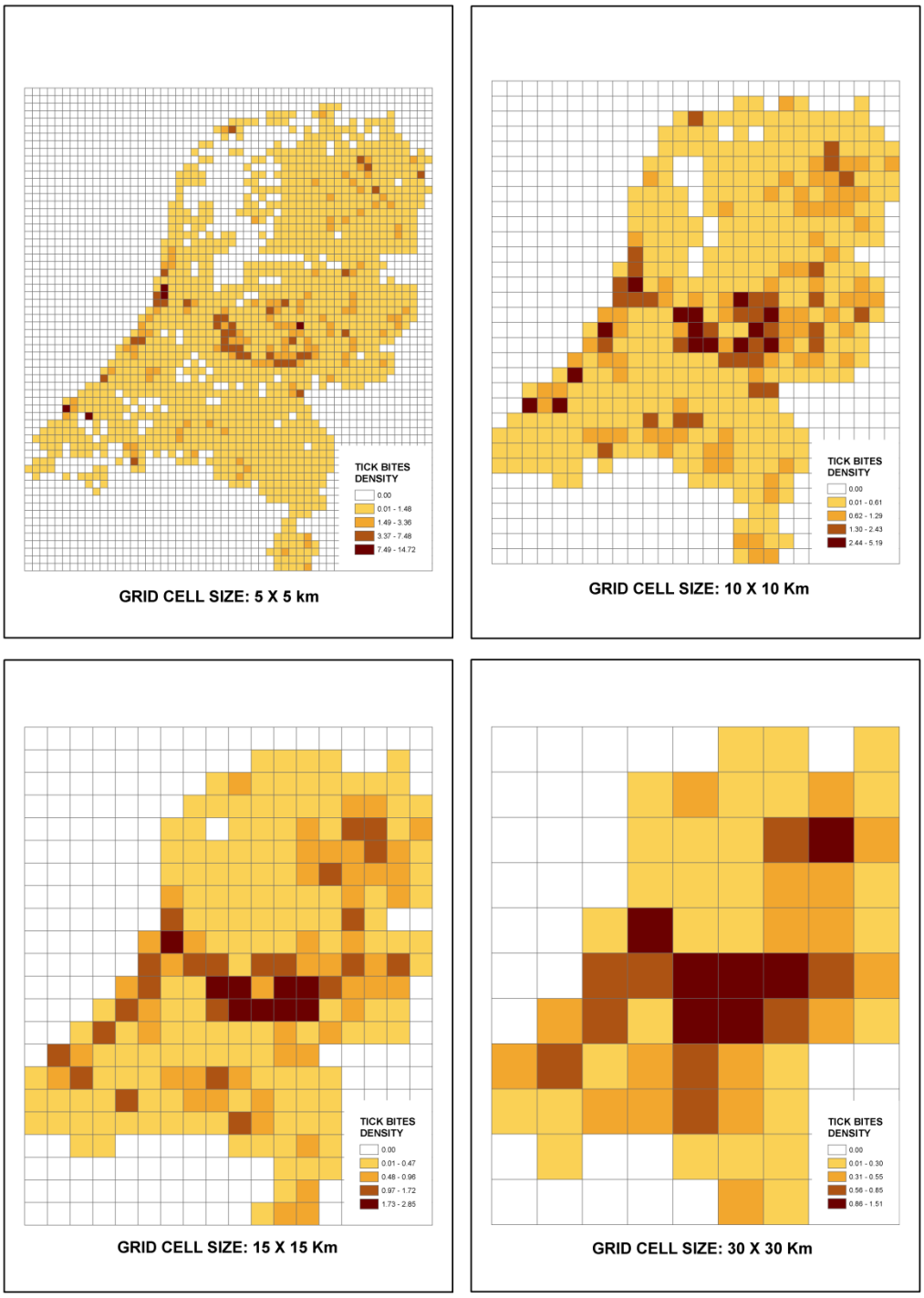


Figure 19: Data aggregated based on several aggregation levels.

The aggregation level which still represents the same behavior of the phenomenon is desired. This is important to ensure that, the result of grouping the individual observations represents the maximum amount of information about the phenomenon (Gozdyra, 2001). As the Netherlands is considered a small country in size; therefore, 30km×30km was too coarse for this analysis. However, after investigating different aggregation levels finer resolution was required. As a result we decided to implement the analysis based on regular grid size 5 km ×5km.

4.2. Data transformation result

The number of reported tick bites varies from year to year in the tick bites dataset. To avoid the big variations in the number of these observations between the years, we attempted to find a similarity measure to describe the behavior of the tick bites occurrence. The main aim of the data transformation process is to convert the tick bite dataset from space in time matrix (S×T) to a temporal variation of the tick bites occurrence across the Netherlands. By this process, we can combine these datasets despite of the variation in the number of observations. The implementation of this step was executed by coding (Appendix D) and a simplified example to explain how this code works was discussed in (section 3.2.2). As a result; the final structure of the dataset consists of 4 columns represent the behavior of the tick bites occurrence in time.

The temporal behavior of the tick bites occurrence varies over the geographic space. In some spatial units the observations was started early and ended early. In some other spatial units the observations were reported in the expected season for this phenomenon between weeks 20 to 30. Moreover, the rest of the spatial units the observations were reported late and continued until the late weeks of the year. In this analysis, all spatial units that have no variations in the temporal behavior were excluded from the dataset. Figure (20) shows a sample of the final structure of the tick bite dataset after applying data transformation step. This data will be used to train SOM in the next stage.

	FID	Shape *	ind25	ind50	ind75	ind100
	65	Polygon	20	25	27	52
	1325	Polygon	22	25	30	51
	387	Polygon	22	25	27	48
	699	Polygon	24	26	35	48
	188	Polygon	26	26	27	47
	689	Polygon	23	24	26	47

Figure 20: Sample of the final structure of ticks bites dataset.

4.3. Data maining and correlation analysis

This section discusses the result of applying the selected data mining technique SOM to the tick bites dataset, besides the results of studying the relationship between several environmental variables and the temporal behavior of the tick bites. The following subsections discuss the training process of SOM, followed by its projection into the geographic space and finally the correlation analysis.

SOM training result

The SOM was trained based on the temporal behavior of the tick bites occurrence in 5224 spatial units of size 5km×5km in The Netherlands. The temporal behavior was selected as a similarity measure between the spatial units. This measure was represented by 4 values (ind25, ind50, ind75, ind100) which are

indicating the week of the year when 25 %, 50%, 75% and 100% of the observations reached. According to this similarity measure SOM was able to group all of the similar spatial units in a cluster.

The SOM training process was started by selecting SOM initialization parameters. Although there is no restrict rule to define these parameters, these parameters were selected based on several literatures, which guide us to set these values. We tried several choices until we achieved proper result. In this research the SOM were initialized based on the following parameters:

- Map dimension: 10×15
- Neighbourhood shape : hexagonal
- Number of iteration: 100000
- Learning rate : 0.05 decrease to 0.01

As a result of the SOM training, the following Figure (21) shows the output plots:

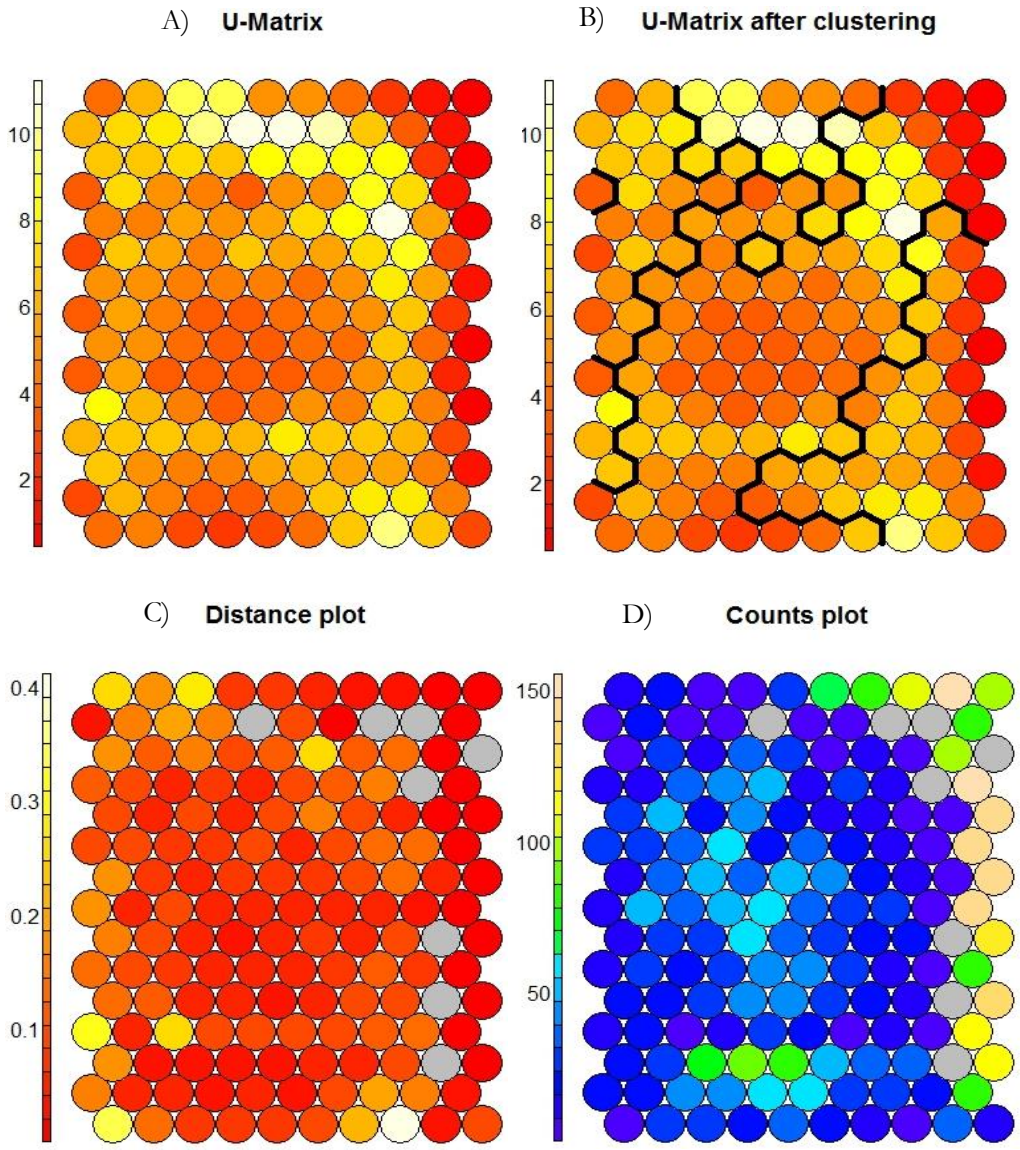


Figure 21: SOM output plots

The U-matrix plot shown in Figure (21-A) represents the Euclidean distance between the codebook vector of each neuron and its neighbors. These distances are represented by different colors; in this case, the red color is corresponding to the neurons that are close to each other in their codebook vectors, while the yellow color indicates the large distance between the codebook vectors. Based on the U-matrix, the analyst can define the number of clusters that may exist in the dataset. The border of the clusters can be identified visually based on the U-matrix, such as what was applied in Tisu(2012). On the other hand, Kohonen package in R provides an option for an automatic clustering process. In this research we applied an automatic clustering approach called hierarchical clustering. Figure (21-B) illustrates the U-matrix after clustering process using hierarchical clustering. The number of clusters was selected to be 4 clusters and the black lines represent the border of each one.

The distance plot displayed in Figure (21-C) is one of the SOM visualizations. This plot represents the mean distance between the units assigned to the neuron and the codebook vector belonging to that neuron. In fact, dark color indicates small distance; this indicates the similarity between objects and the best matching unit. In contrast, the light color indicates large distance between the object and the best matching unit. Here, dark color assigned to most of the neurons, this indicates good result in this case.

The last visualization appears in Figure (21-D) is called counts plot. This visualization indicates the number of data objects mapped to each neuron. The light color means a large number of data objects were mapped to that neuron while the dark color indicates a small number of data objects. Moreover, no data objects were assigned to those neurons depicted in gray. The edge effect is obvious in our case here where the neurons with a large number of data objects are clustering on the edge.

SOM clustering

Clustering process in this research performed automatically using hierarchical clustering in Kohonen package. Four clusters were defined as illustrated in Figure (22). A few neurons in this figure appears isolated as a separate cluster, were investigated these neurons individually, and we found out that they are belonging to that cluster embedding them. This might be caused by error of hierarchical clustering in Kohonen package.

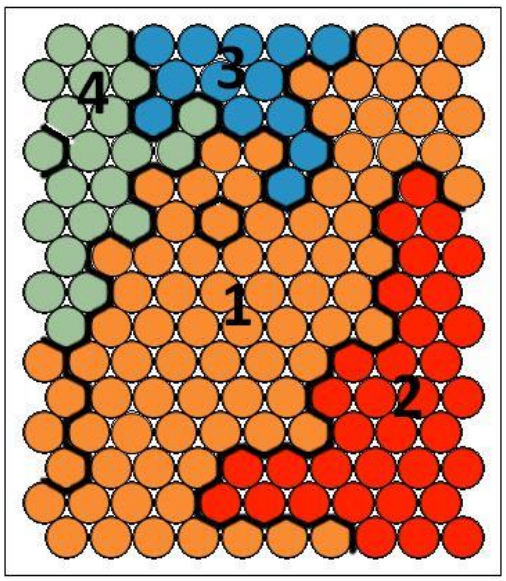


Figure 22: SOM clusters

The spatial units belonging to the same cluster are similar to each other in their temporal behavior. For further understanding the codebook vectors of each cluster was plotted to display the temporal behavior in these clusters. Figure (23) shows the temporal behavior of tick bites occurrence in each cluster: in the spatial units belonging to cluster 1, the tick bites were reported during the expected season for ticks. While the tick bites occurrence was fast (started early and finished early) in the spatial units those were assigned to cluster 2. On the contrary, cluster 3 represents the spatial units where the reporting tick bites was continued for a longer period (starts early and continued to late weeks of the year).In addition, cluster 4 represents late tick bites occurrence.

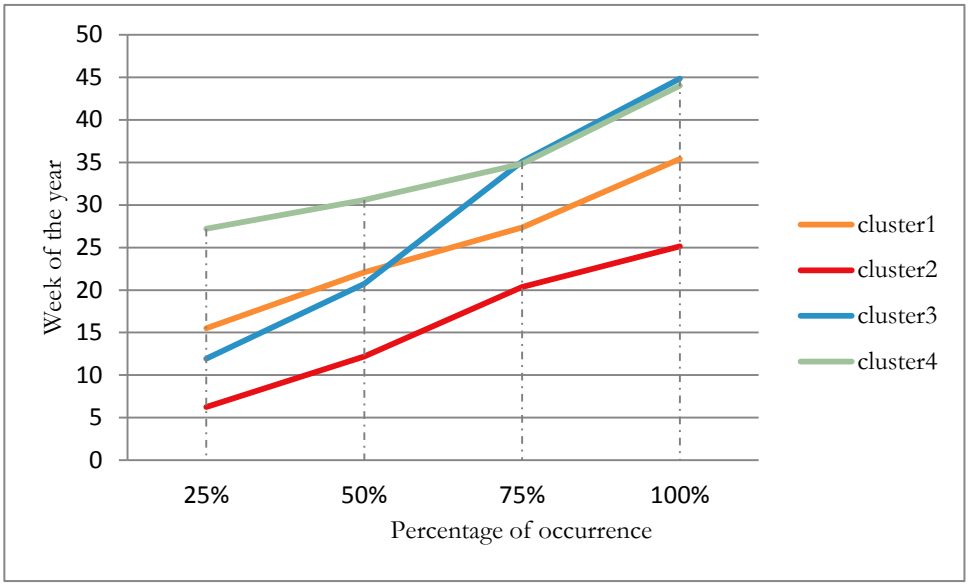


Figure 23: Average of de normalized codebook vector

The total number of the input dataset used to train SOM was 5224 observations represent the temporal behavior of tick bite occurrence in the spatial units. Table (9): Percentage of the input data per cluster the percentages of the input data in each cluster per year. From the values shown in this table, we can observe that the data are unevenly distributed among the clusters. Actually, cluster 1 was the dominant cluster and has the highest percentage of the data every year as shown in Figure (24). Cluster 2 came in the second rank this was the case for every year except year 2007 which shows the data percentage of cluster 2 (early tick bites occurrence), was more than data percentage in cluster 4 (late tick bites occurrence).

Table (9): Percentage of the input data per cluster (2006-2013)

Cluster	Year2006	Year2007	Year2008	Year2009	Year2010	Year2011	Year2012	Year2013
1	43.16	28.07	33.42	56.57	40.95	40.83	53.29	55.94
2	15.15	26.97	18.45	12.66	16.96	17.50	13.54	9.52
3	10.10	25.32	23.94	11.85	19.68	20.42	14.08	13.22
4	31.58	19.63	24.19	18.91	22.39	21.25	19.07	21.32

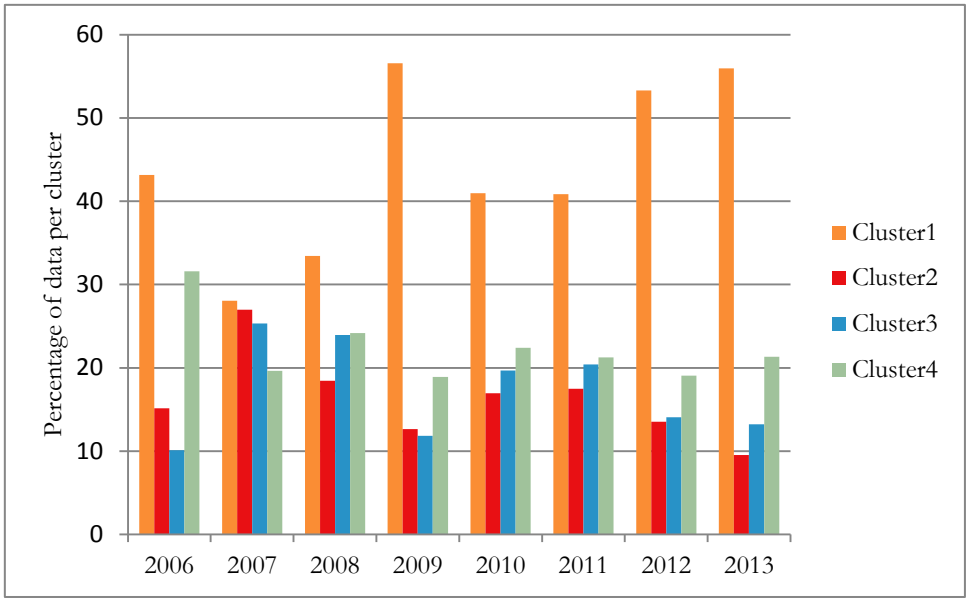


Figure 24: Percentage of the input data per cluster

SOM result projection into the geographic space

For further explorations, the SOM results were separated yearly and projected into the geographic space. The following Figures (25) and (26), illustrate the result of projecting SOM clusters into the geographic space, each spatial unit depicted in the same color of its corresponding cluster.

The results were visually interpreted; we recognized that cluster 1 was the dominant cluster every year. However, the clusters are randomly distributed in the geographic space and there was no cluster absolutely dominant in a certain part of the country. Visually, there was no particular pattern in the clusters distribution. Therefore, stability check was performed to assess the stability of the spatial units within the same clusters over the years. The result of stability check showed that all the spatial units were not stable in a certain cluster during our study temporal span and they were changing from a cluster to another yearly.

The absence of the pattern and the variability of the temporal behavior of tick bites occurrence in the spatial unit over time could be due to the following factors:

- The variability of the environmental conditions under which the tick bites occurred over space and time.
- The variability of human recreational activities space and time.

In the next section we will study the relationship between the selected environmental variables and the corresponding cluster. To find out if any environmental variables have clear impact on the behavior of the tick bites phenomenon.

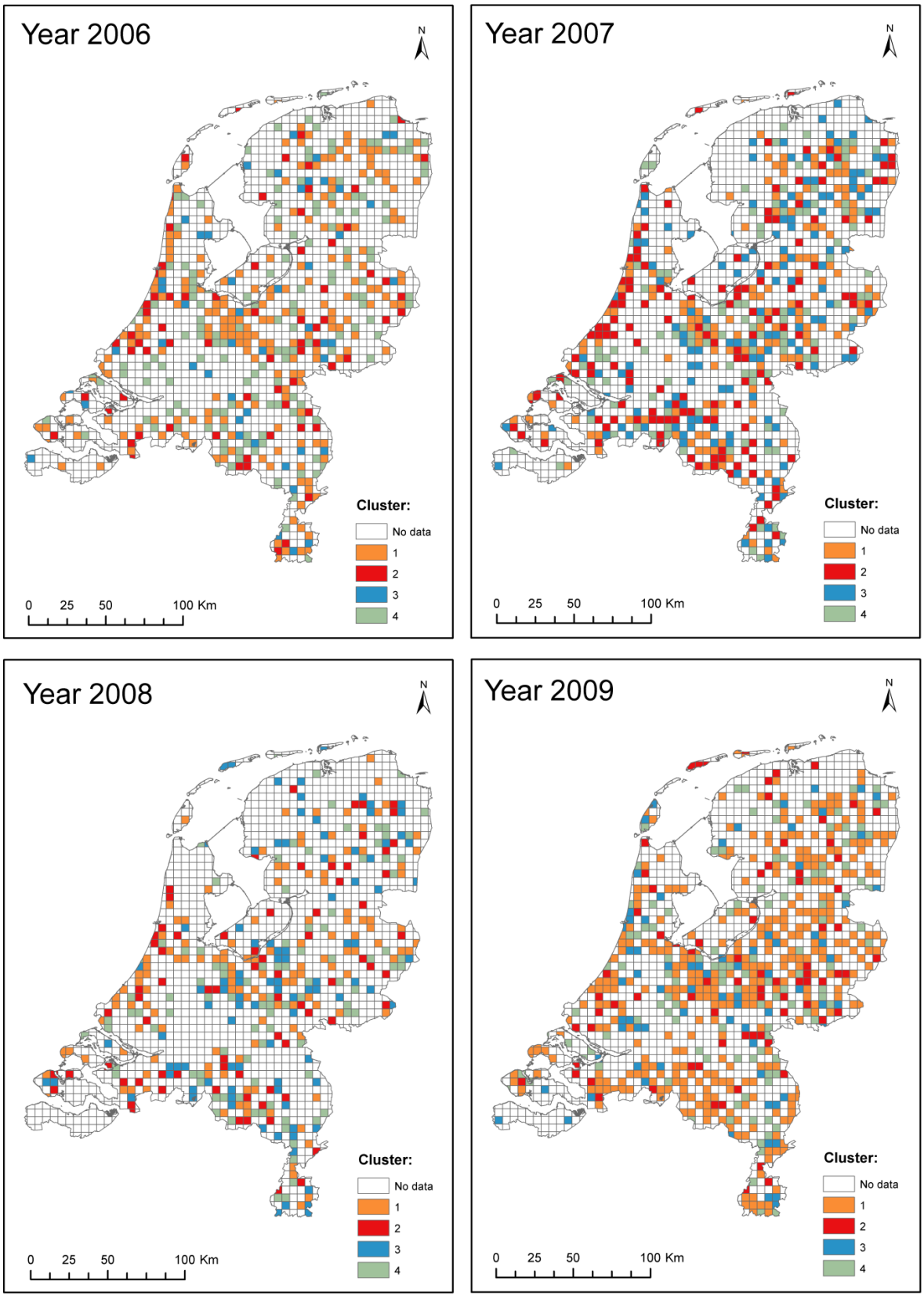


Figure 25: Projection of SOM clusters into the geographic space (2006-2009)

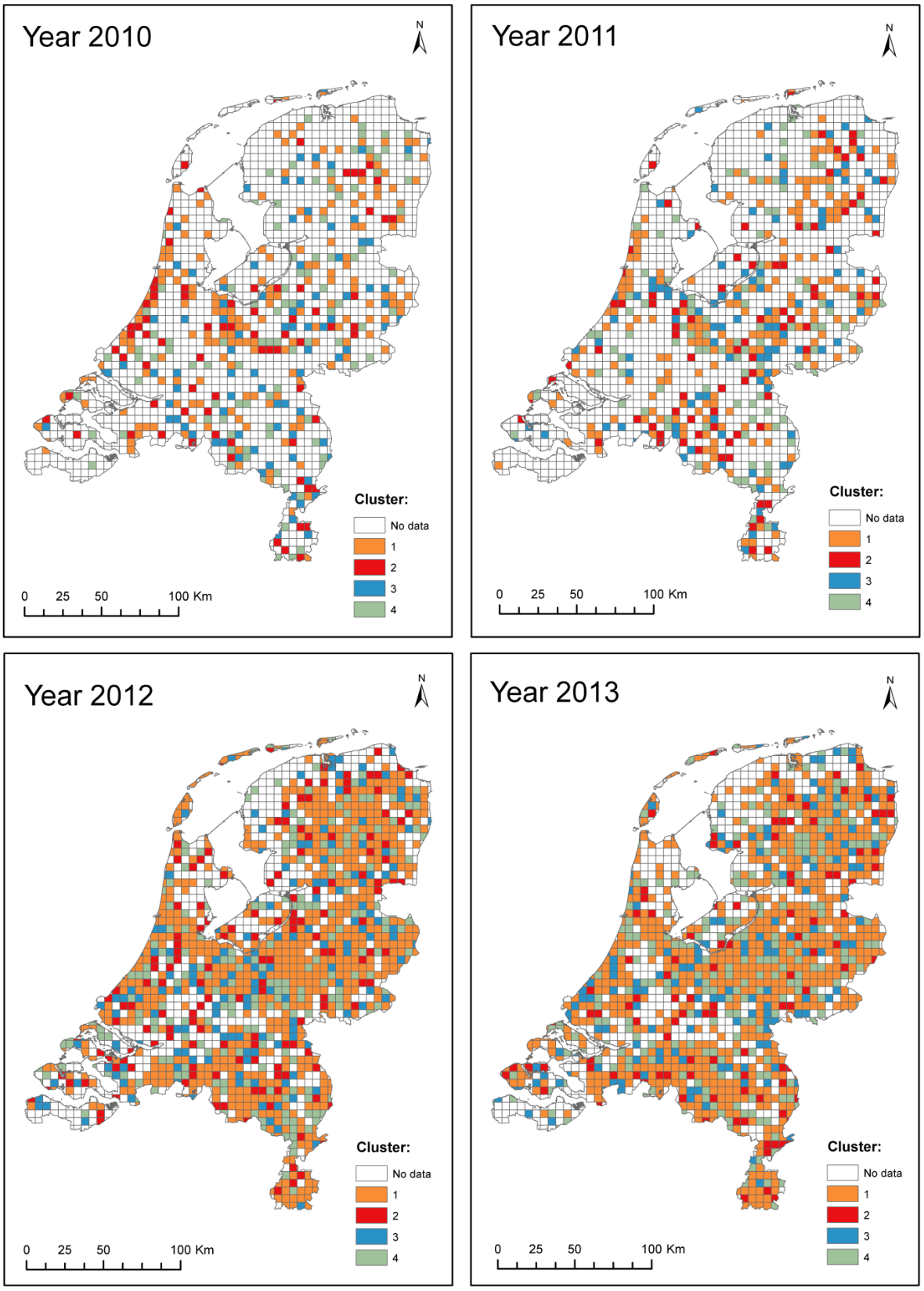


Figure 26: Projection of SOM clusters into the geographic space (2010-2013)

Correlation analysis between clusters and Environmental variables

This step of the analysis attempted to study the relationship between the environmental variables and the clusters to find out if any of these explanatory factors involved in this analysis has a clear impact on the temporal behavior of the tick bites occurrence, and can be used to predict the phenomenon. The following subsection explain the basic of selection the explanatory variable and the result of the analysis.

Explanatory variables selection

The research work started with a literature review, to understand the theoretical background of the research in terms of the ecological and technical problem. As a result, this phase has revealed on the environmental variables that have impact on ticks' abundance and relevant to this research. The selection of the environmental variable was based on three aspects:

- The key environmental variables that have an impact on tick distribution which defined from Literature review (section2.3).
- The knowledge that tick bite occurrence and be modelled by :

  (Tick bite = human activity × tick activity)
- The availability of the data.

As a result the following variables were selected to be included in the analysis:

- Land use and cover percentage in each spatial unit.
- Forest mean patch area as indicator of forest fragmentation.
- Temperature (weekly).
- School vacation during the year.

Correlation analysis:

This subsection discusses the result of analysing the relationship between the explanatory variables and the clusters as the following:

LULC percentage: Initially, the correlations between the clusters and the LULC percentages were analysed, to figure out if there is a significant variation between the LULC among the clusters. To implement this step, the data previously prepared by calculating the percentage of several LULC types were calculated in each spatial unit as discussed in section (3.1.2). Further, this dataset was joined with the result of SOM clusters, and statistic was performed.

The average percentages of the following LULC types were computed in each cluster. Specifically, recreation, forest, agriculture, water and built up regions. The results of these statistics are shown in Figure (27). This Figure illustrates the average percentage of each LULC type in all spatial units belonging to each cluster. There was no big variation between the clusters in any LULC type. For instance, in recreation areas all clusters are relatively the same. In the other LULC types, the difference between the clusters was minor. For example, in the forest, cluster 2 (early tick bite occurrence) has a relatively higher average percentage than other clusters. Despite of these slight differences, still the variation is not clear. One cannot see that any of the clusters were dominant in specific LULC. This result implies that LULC types have no clear impact on the temporal behavior of tick bite occurrence.

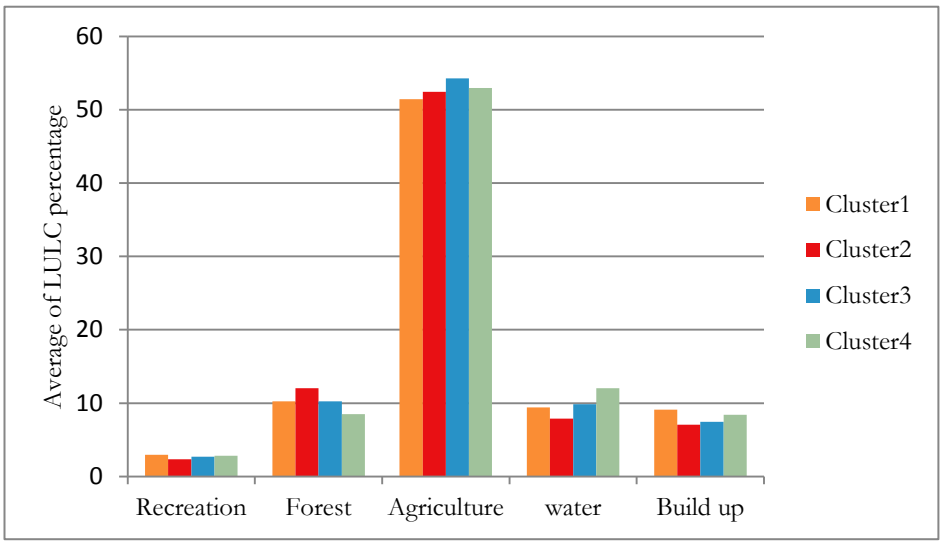


Figure 27: Average of LULC percentage per cluster

Forest fragmentation: this is considered as one of the factors that may have an impact on tick abundance. Forest fragmentation represents a form of habitat fragmentation, that occurs when continuous forest is fragmented in to several small patches either by a natural process (e.g. fire or climate) or human activities (e.g. urbanization, cleaning for agriculture). This, leads to a biological and physical changes in the forest environment (Jha et al., 2005). In this context, we address the forest fragmentation as an environmental variable, to assess, whether the forest fragmentation has a strong impact on the temporal behavior of the tick bites occurrence.

Forest mean patch area was used to approximately assess the level of forest fragmentation in each spatial unit. Further, the average of mean patch area was computed in each cluster and the result shown in Figure (28). In fact, the result doesn't show any specific pattern during all years. Therefore, there was no clear impact of the forest fragmentation on the temporal behavior of tick bite occurrence.

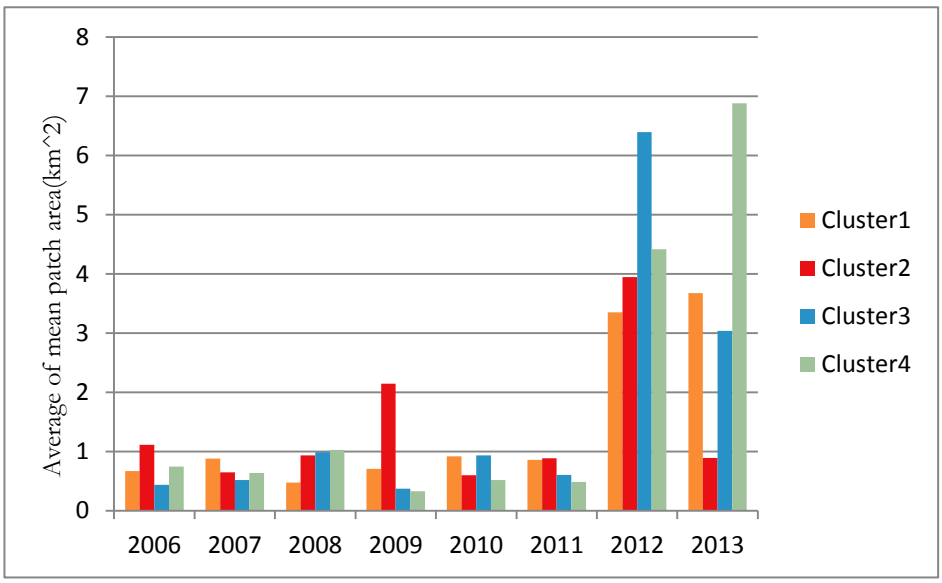


Figure 28: Average of forest mean patch area per cluster

Temperature: climate factor such as temperature has an impact on tick abundance. Also, human recreation activities could be related to the temperature. For instance, during the summer more recreation activities are expected than in winter. Therefore, maximum mean temperature is included in this analysis as an explanatory variable, to study the correlation between the temperature and the clusters. To find out if the temperature has impact on the temporal behaviour of the tick bites occurrence.

According to the completeness of the temperature dataset, we prepared in the early stage of this research (section 3.1.2) we select the warmest year (2007) and the coldest year (2009) to implement the analysis. These datasets represent the mean of maximum temperature in each spatial unit every week. Then the averages of the weekly temperature per cluster were computed and plotted as shown in Figure (29) and (30). The differences in the temperatures between the clusters were insignificant and cannot be recognized from the figure, this implies there is no variation in the temperatures between the clusters.

Figure 29: Average weekly temperature per cluster (Year2007)

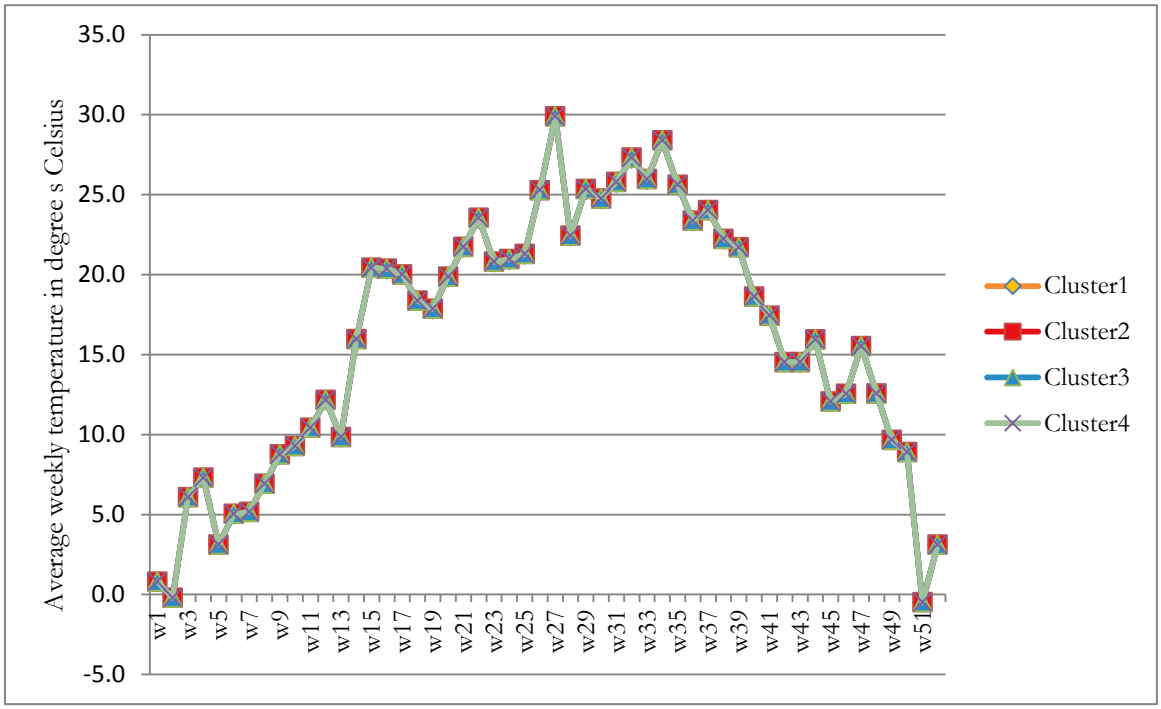


Figure 30: Average weekly temperature per cluster (Year 2009)

School vacation: In fact, increase in recreation activities in nature are expected during the vacations. This may raise the chance for human exposure to a tick bite. School vacation dataset was collected and prepared in a way to be suitable to this analysis. From the histograms' analysis previously shown in Figure (17) and (18) in section (4.1.3) the autumn peak was coincided with the school autumn vacation. For further analyses, we join the school vacation dataset with the result of SOM, then the spatial units that have week 42, 43and 44 were analysed, to check if there is a correlation between the temporal behavior tick bites occurrence and the school vacations. The result shown in Figure (31) , from this chart one can recognize that every year cluster 4 has the second rank, meaning, there were a significant number of spatial units belonging to cluster 4 (late tick bite occurrence). This indicates there is a relation between the school vacations, and the occurrence of tick bites.

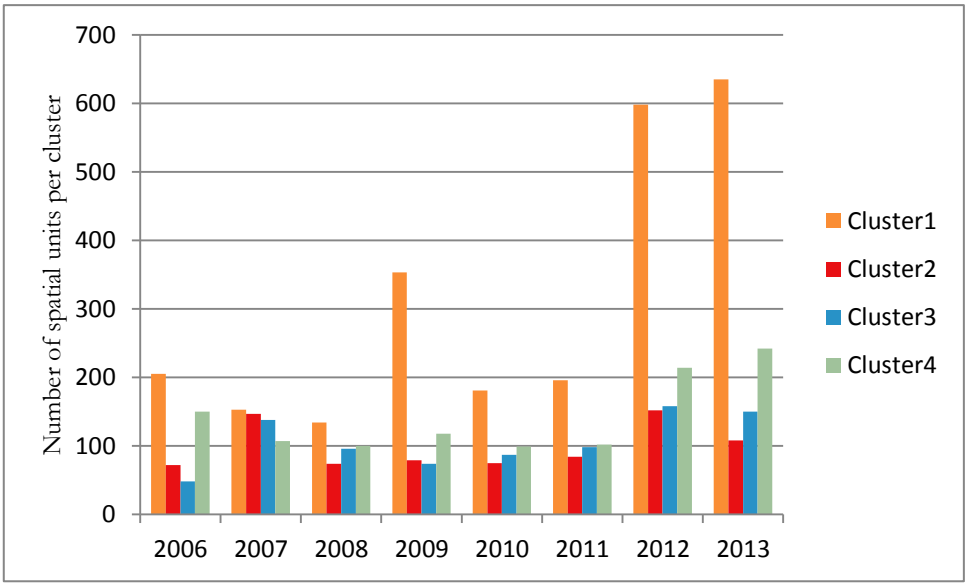


Figure 31: Number of spatial units that have autumn school vacation per cluster

Discussion:

The SOMs results were mapped onto the geographic space, the visual inspection of this projection did not show any particular spatial pattern in the distribution of the clusters in the Netherlands. In addition, the result of correlation's analysis between the clusters and the selected environmental variables did not show any clear variations in the clusters, except the school autumn vacation that indicates there is a relationship between the school autumn vacation, and the late occurrence of the tick bites.

In fact, these results could be as a consequent of one or more of the following reasons: First, the SOM parameters were not the optimal for tick bites dataset. Second, aggregation level was not the correct choice. Third, the quality, heterogeneity of the VGI data (tick bites dataset), and the differences in the number of the observations were biasing the output. Fourth, the impacts of the environmental variables vary over time and space. Fifth, the human activity has a strong impact in the tick bites occurrence. The following paragraphs will address these possibilities one by one:

The first possible reason, that the SOM parameters were not the best choice. In fact, there is no particular rule to select the SOM parameters but in this research, we followed some rules were recommended in literature and we tried several options. Finally, the parameters those defined the cluster clearly were selected. Therefore, this reason may have a minor effect but not the main reason behind the absence of the pattern.

The second possible reason, that the selected spatial aggregation level was not the proper choice. Actually, the aggregation level was not selected arbitrary, the code that we prepared to implement the data aggregation task allows for the analyst to try several aggregation levels. Therefore, several aggregations were investigated and the one that represents the phenomenon was selected. From our point of view, the Netherlands is small country in size and the selected aggregation level (5km×5km) was a proper choice. For further confirmation, we run the SOM with same parameters based data aggregated on several aggregation levels such as (10km×10km) and (30km×30km), and the result did not show any specific pattern. Sample of the SOM result based on the mentioned aggregation levels is provided in appendix (E).

The third possible reason related to the quality, heterogeneity of the VGI datasets and the variation in the number of observations reported yearly. The lack in VGI data quality is a common problem. Therefore, the data preparation stage was started with quality check and filtration process for the observations that are erroneous. According to the variations in the number of observations between the years, this was avoided by constructing the similarity measure representing the temporal behavior of the tick bite occurrence. Moreover the SOM was trained based on the temporal behavior of the tick bites occurrence in each spatial unit and not based on the number of the observations. In fact, our method is sensitive to the temporal accuracy of the VGI dataset, because the similarity measure was derived based on the data after temporal and spatial aggregation. Consequently, the tick bites occurrence date and its location have an influence on the results of this analysis.

The fourth possible reason, that the impact of the environmental variables. The environmental conditions vary in space and time. Regardless of this fact, the correlations results didn't show significant variations of these environmental variables between the clusters. Except the school vacations which indicated that there is a relationship between the school vacations and the late occurrence of tick bites. This result implies that no one of the other environmental variables can be used to predict the temporal behavior of the tick bites.

The fifth possible reason is the impact of human activities. In reality, the tick bites occurrence is highly correlated with the human activities. Human activities are varying over space and time, and in my opinion, it has the major role in the result of our analysis. The absence of the pattern is influenced by the variations in human activities. Specifically, the tick bite occurrence is correlated with the human recreational pressure. One can find a spatial unit with a significant number of tick bites although it has a significant percentage of water (where is no ticks), but there is small part as recreation area, which leads to highlight this spatial unit as high risk. Moreover, the spatial units that have a high number of tick bites were inspected visually; we notice that these spatial units have a similarity in the LULC type. There is a common combination of three LULC types (forest, recreation, and built up areas) exist in these units. As shown in Figure (32), both of these units have the same LULC. But more tick bites were reported, when these three types of LULC close to each other or sharing boundaries. It is clear that when recreation areas are near by a forest this increase the chance of human exposure to get a tick bite.

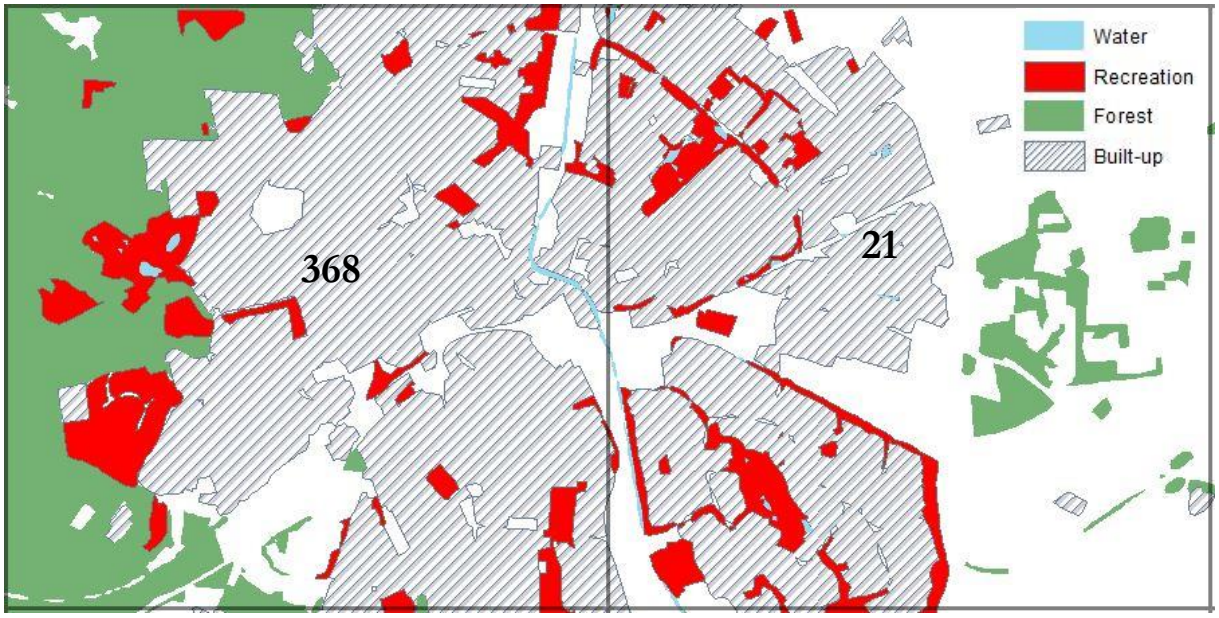


Figure 32: Visual inspection of two spatial units.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

The main objective of this research was studying the correlation between the tick bite occurrences and several environmental variables. For this we mined a VGI dataset with tick bites occurrence records as reported by volunteers in the Netherlands.

To achieve the main objective of this research, analytical workflow was implemented. The workflow consists of three main stages:

1) Data preparation and aggregation: at this stage the quality of the tick bites dataset, was verified and its observations were aggregated in space and time. Moreover, all other relevant datasets such as LULC, temperature, forest fragmentation and school's vacations datasets were prepared.

2) Data transformation stage: this stage was applied only for the complete tick bite datasets which were heterogeneous in its source and contents. Specifically, the VGI datasets (tick bites dataset) were consisting of spatio-temporal observations and there were significant differences in the numbers of the observations between the years. In order to avoid these differences, we designed computational steps to construct the temporal behavior of the tick bites occurrence which was used later as a similarity measure to cluster the spatial units.

3) Data mining and correlation analysis: in this stage the complete tick bite dataset were used to train SOM; the SOM could cluster the spatial units based on their similarity in the temporal behavior of the tick bites occurrence. Further, the correlations between the selected environmental variables and clusters were analysed in this stage.

By the implementation of the research workflow, we achieved the main objective of this research. The following section discusses the answers for the research questions:

1) **What are the key environmental variables that affect the tick abundance addressed in literature?**

The literature review stage was revealed the key environmental variables that have an influence on ticks abundance in nature. Ticks are unevenly distributed in nature, and their distribution influenced by several variables such as: climate variables, land cover, landscape structure and host availability.

Several climate factors have an impact on ticks' abundance (e.g. temperature, humidity, precipitation, water saturation…etc.).Mainly, ticks activity and development are regulated by the temperature and humidity. Ticks need a blood meal to survive and develop from one life stage to another. Therefore, they start hold on the vegetation seeking for a host this is called questing activity. Questing behavior is sensitive to the temperature, several studies shown that ticks start questing when the daily maximum temperature exceed 7°C (Sprong et al., 2012). Moreover, the water balance in ticks is influenced by the relative humidity, when ticks are questing on top of the vegetation leaves or on the grass, it may lose its water content. Bennet et al., (2006) stated that, when the relative humidity above 86% ticks can actively seek for a host.

In nature, forest and vegetated areas are suitable habitat for ticks more than pastures.in fact; vegetation has two roles the first, in maintaining the conditions of the microclimate surrounding the ticks. The second is providing questing sites for the ticks while waiting for a host. NDVI has been linked with tick abundance and tick borne disease transmission by (Kalluri et al., 2007).

Previous studies addressed forest fragmentation as one of the factors that has an impact on tick abundance. Forest fragmentation leads to increase the number of tick host and increase the density of rodents, which leads to increase the spatial distribution on the ticks.

2) What aggregation level is suitable for the analysis of the tick bites?

To answer this question, Python code was prepared to allow us to try several aggregation levels. The study area was divided to regular equal size sub regions. The Netherland is a small country; therefore, 5 km×5 km was selected as a proper aggregation level. That was the aggregation level represents the original behavior of the phenomenon as shown in Figure (19). However, our research workflow is adaptable for any spatio-temporal dataset related to any country, for this reason, if we applied our workflow based on European dataset then another aggregation level could be required. According to the temporal aggregation level we, select the weekly level, since monthly is too coarse.

3) Which exploratory techniques highlight the period with a higher occurrence tick bites?

Frequency histogram was selected as an exploratory analysis tool. Frequency histogram provides visual display of the data. The tick bites dataset consist of observation have reference in space (absolute x, y coordinates) and in time (bite's occurrence date). Initially, the week of the year WOY was generated from bite's occurrence dates for all observations in the dataset. Then frequency histograms used to visualize the number of observations reported weekly. The dataset was separated per year, and displayed using histograms to highlight the high risk period every year. Two peaks appear in the histograms: the main peak in between weeks 20 and 30 and autumn peak around week 42, 43.

4) How to handle the problem of VGI data quality?

The implementation of the workflow was started by the quality verification step, in order to handle the quality problems of the VGI dataset (tick bite dataset). Initially, the original dataset mapped into geographic space to understand the spatial distribution of the observations in the Netherlands and to check the quality of these observations. Several explorations were performed using attribute and spatial queries. Then, all the observations that had an obvious erroneous value in their location or attributes were filtered out. For instance, any observation has incomplete attributes (specifically, location) were removed. In addition, the observations reported out of our research temporal span were removed; all observations were reported outside the Netherlands or in the water, were removed. According to the date of each observation, we assumed it is correct since we can't be sure of its accuracy.

5) What are the most appropriate parameters for the SOM?

The Data mining stage in the workflow includes selecting the SOM parameters and train the SOM based on the complete tick bite dataset. In fact, there are no restricting rules to select SOM parameters, in this research we followed the guide rules were recommended by previous researcher. Besides, several options were examined until and the following settings were selected:

- Map dimension: 10×15
- Neighbourhood shape : hexagonal
- Number of iteration: 100000
- Learning rate : 0.05 decreases to 0.01

6) Which environmental variables have more impact on the phenomenon of tick bites occurrence?

The result of projecting the clusters derived from the SOM onto the geographic space did not show any particular pattern. And according to the correlation analysis between the clusters and the selected environmental variables there were no significant variations in the environmental variables between the clusters. This implies no one of these variables has strong impact on the phenomenon, and cannot be used to predict the temporal behavior of the tick bites occurrence in the spatial units defined in this study.

5.2. Recommendations

At the end of this research, this section provides several recommendations for further analysis and future work.

- Although there are several reasons behind the absence of the spatial patterns in the dataset (at spatial aggregation level (5km×5km) and temporal aggregation: weekly), we highly recommend applying the same method based on different aggregation levels or using dataset with different characteristics. For instance, further analysis based on "Tekenradar" dataset only, may lead to a better result. Actually, mixing the tick bites observations from different source "Tekenradar "and "Natuurkalender" was not an adequate choice for our analysis: because the attributes of the dataset collected from "Natuurkalender" has only one date that was considered as bite's occurrence date, and the method we applied was sensitive to the temporal accuracy of the VGI dataset, this could influence the results.

- The implemented workflow is adaptable to any spatio-temporal dataset it could be applied for any other dataset. In fact, if it is applied on a larger level for instance European dataset the result may lead to a particular pattern due to the significant variation in the environmental variables between the countries rather than in the case on the Netherlands. Here, the analyst needs to select a coarser aggregation level.

- Trying other relevant explanatory variables. This case study is highly correlated to human activities such as recreation. Involving recreation data such as the number of national parks visitors may show a clear relationship.

- To reduce the effect of the VGI dataset quality, define a temporal threshold for selected observations, this implies removal all the early and late observations since these observations have higher uncertainty in time.

- Monitoring network such as "Tekenradar" needs to apply some rules to enhance and control the quality of the VGI dataset. For instance, by limiting the reporting date to be within one week since the incident occurs so that volunteers can't report approximate date for a tick bite occurred in the past. Moreover. if they by mistake enter reporting date before the biting date an error message should appear on the screen ask the volunteer to fix it.

LIST OF REFERENCES

Al-Ahmadi, K., & Al-Ahmadi, S. (2013). Rainfall-Altitude Relationship in Saudi Arabia. *Advances in Meteorology*, *2013*, 1–14. doi:10.1155/2013/363029

Andrienko, G., Malerba, D., May, M., & Teisseire, M. (2006). Mining spatio-temporal data. *Journal of Intelligent Information Systems*, *27*(3), 187–190. doi:10.1007/s10844-006-9949-3

Andrienko, Natalia; Andrienko, G. (2006). *Exploratory Analysis of Spatial and Temporal Data : A Systematic Approach*. Dordrecht: Springer-Verlag Berlin and Heidelberg GmbH.

Augustijn, E.-W., & Zurita-Milla, R. (2013). Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. *International Journal of Health Geographics*, *12*(1), 60. doi:10.1186/1476-072X-12-60

Baneth, G. (2014). Tick-borne infections of animals and humans: a common ground. *International Journal for Parasitology*, *44*(9), 591–596. doi:10.1016/j.ijpara.2014.03.011

Barrios, J. M., Verstraeten, W. W., Maes, P., Aerts, J. M., Farifteh, J., & Coppin, P. (2013). Seasonal vegetation variables and their impact on the spatio-temporal patterns of nephropathia epidemica and Lyme borreliosis in Belgium. *Applied Geography*, *45*, 230–240. doi:10.1016/j.apgeog.2013.09.019

Barrios, J. M., Verstraeten, W. W., Maes, P., Clement, J., Aerts, J. M., Farifteh, J., Coppin, P. (2012). Remotely sensed vegetation moisture as explanatory variable of Lyme borreliosis incidence. *International Journal of Applied Earth Observation and Geoinformation*, *18*, 1–12. doi:10.1016/j.jag.2012.01.023

Beaujean, D. J. M. A., Gassner, F., Wong, A., Steenbergen van, J. E., Crutzen, R., & Ruwaard, D. (2013). Determinants and protective behaviours regarding tick bites among school children in the Netherlands: a cross-sectional study. *BMC Public Health*, *13*, 1148. doi:10.1186/1471-2458-13-1148

Bennet, L., Halling, A., & Berglund, J. (2006). Increased incidence of Lyme borreliosis in southern Sweden following mild winters and during warm, humid summers. *European Journal of Clinical Microbiology & Infectious Diseases : Official Publication of the European Society of Clinical Microbiology*, *25*(7), 426–32. doi:10.1007/s10096-006-0167-2

Bogorny, V., & Shekhar, S. (2010). Spatial and Spatio-temporal Data Mining. In *2010 IEEE International Conference on Data Mining* (pp. 1217–1217). IEEE. doi:10.1109/ICDM.2010.166

Boyard, C., Barnouin, J., Gasqui, P., & Vourc'h, G. (2007). Local environmental factors characterizing Ixodes ricinus nymph abundance in grazed permanent pastures for cattle. *Parasitology*, *134*(Pt 7), 987–94. doi:10.1017/S0031182007002351

Brownstein, J. S., Holford, T. R., & Fish, D. (2003). A climate-based model predicts the spatial distribution of the Lyme disease vector Ixodes scapularis in the United States. *Environ Health Perspect*, *111(9)*, 1152–1157.

Brunsdon, C., Fortheringham;, A. S., & E.Charlton, M. (1996). Geographically Weighted Regression: AMethod for Exploring Spatial Nonstationarity. *Geographical Analysis*, *28*(4). Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1538-4632.1996.tb00936.x/pdf

Chon, T.-S. (2011). Self-Organizing Maps applied to ecological sciences. *Ecological Informatics*, *6*(1), 50–61. doi:10.1016/j.ecoinf.2010.11.002

Coons, L. B., & Rothschild, M. (2004). *Encyclopedia of Entomology* (pp. 2240–2262). Dordrecht: Kluwer Academic Publishers. doi:10.1007/0-306-48380-7

Dale, M. R. T., & Fortin, M.-J. (2009). Spatial autocorrelation and statistical tests: Some solutions. *Journal of Agricultural, Biological, and Environmental Statistics*, *14*(2), 188–206. doi:10.1198/jabes.2009.0012

Dantas-Torres, F., Chomel, B. B., & Otranto, D. (2012). Ticks and tick-borne diseases: a One Health perspective. *Trends in Parasitology*, *28*(10), 437–46. doi:10.1016/j.pt.2012.07.003

Den Boon, S., JF, S., LM, S., AW, S., B, D. van L., & W, van P. (2004). Doubling of the number of cases of tick bites and lyme borreliosis seen by general practitioners in the Netherlands. *Ned Tijdschr Geneeskd*, 665–70. Retrieved from http://bvsalud.org/portal/resource/en/mdl-15106318

Doelen Tekenradar.nl - Over Tekenradar.nl - Home - Tekenradar. (2014). Retrieved August 20, 2014, from https://www.tekenradar.nl/over-tekenradar-nl/doelen-tekenradar-nl

Dube, J. ;, & Legros, D. (2014). *Spatial Econometrics using Microdata*. Retrieved from https://books.google.nl/books

Ejarque-Gonzalez, E., & Butturini, A. (2014). Self-organising maps and correlation analysis as a tool to explore patterns in excitation-emission matrix data sets and to discriminate dissolved organic matter fluorescence components. *PloS One*, *9*(6), e99618. doi:10.1371/journal.pone.0099618

Estrada-Peña, A. (2001). Forecasting habitat suitability for ticks and prevention of tick-borne diseases. *Veterinary Parasitology*, *98*(1-3), 111–132. doi:10.1016/S0304-4017(01)00426-5

Estrada-Peña, A., & de la Fuente, J. (2014). The ecology of ticks and epidemiology of tick-borne viral diseases. *Antiviral Research*, *108C*, 104–128. doi:10.1016/j.antiviral.2014.05.016

Estrada-Peña, A., Gray, J. S., Kahl, O., Lane, R. S., & Nijhof, A. M. (2013). Research on the ecology of ticks and tick-borne pathogens--methodological principles and caveats. *Frontiers in Cellular and Infection Microbiology*, *3*, 29. doi:10.3389/fcimb.2013.00029

Estrada-Peña, A., Martinez, J. M., Sanchez Acedo, C., Quilez, J., & Del Cacho, E. (2004). Phenology of the tick, Ixodes ricinus, in its southern distribution range (central Spain). *Medical and Veterinary Entomology*, *18*(4), 387–97. doi:10.1111/j.0269-283X.2004.00523.x

Estrada-Peña, JM, V., & C., S. A. (2006). The tick Ixodes ricinus: distribution and climate preferences in the western Palaearctic. *Med Vet Entomol*, *20(2)*, 189–97. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16874918

Fix, M. (2006). *Handbook of Bioterrorism and Disaster Medicine*. (R. E. Antosia & J. D. Cahill, Eds.) (pp. 299–305). Boston, MA: Springer US. doi:10.1007/978-0-387-32804-1

Fritz, C. L. (2009). Emerging tick-borne diseases. *The Veterinary Clinics of North America. Small Animal Practice*, *39*(2), 265–78. doi:10.1016/j.cvsm.2008.10.019

Gilbert, L. (2010). Altitudinal patterns of tick and host abundance: a potential role for climate change in regulating tick-borne diseases? *Oecologia*, *162*(1), 217–25. doi:10.1007/s00442-009-1430-x

Giraudel, J. L., & Lek, S. (2001). A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling*, *146*(1-3), 329–339. doi:10.1016/S0304-3800(01)00324-6

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 211–221. doi:DOI 10.1007/s10708-007-9111-y

Goodchild, M. F. (2009). Geographic information systems and science: today and tomorrow. *Procedia Earth and Planetary Science*, *1*(1), 1037–1043. doi:10.1016/j.proeps.2009.09.160

Goodchild, M. F., & Li, L. (2012a). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. doi:10.1016/j.spasta.2012.03.002

Goodchild, M. F., & Li, L. (2012b). Assuring the quality of volunteered geographic information. *Spatial Statistics*, *1*, 110–120. doi:10.1016/j.spasta.2012.03.002

Gozdyra, P. (2001). *DATA SPATIAL AGGREGATION ISSUES IN PUBLIC HEALTH ANALYSE*. University of Toronto. Retrieved from https://tspace.library.utoronto.ca/bitstream/1807/15915/1/MQ63052.pdf

Gray, J. S., Dautel, H., Estrada-Peña, A., Kahl, O., & Lindgren, E. (2009). Effects of climate change on ticks and tick-borne diseases in europe. *Interdisciplinary Perspectives on Infectious Diseases*, *2009*, 593232. doi:10.1155/2009/593232

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, *37*(4), 682–703. doi:10.1068/b35097

Himberg, J., Ahola, J., Alhoniemi, E., Vesanto, J., & Simula, O. (2001). The Self-Organizing Map as a Tool in Knowledge Engineering. *World Scientific*, *2*, 35–65. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.1.6243

Hochachka, W. M., Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., & Kelling, S. (2007). Data-Mining Discovery of Pattern and Process in Ecological Systems. *Journal of Wildlife Management*, *71*(7), 2427. doi:10.2193/2006-503

Hofhuis, A., van der Giessen, J. W. B., Borgsteede, F. H. M., Wielinga, P. R., Notermans, D. W., & van Pelt, W. (2006). Lyme borreliosis in the Netherlands: strong increase in GP consultations and hospital admissions in past 10 years. *Euro Surveillance : Bulletin Européen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*, *11*(6), E060622.2. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16819128

How do humans get Lyme disease? (2014). Retrieved September 09, 2014, from http://www.patient.co.uk/health/lyme-disease-leaflet#

How ticks spread disease. (2014). Retrieved September 10, 2014, from http://www.cdc.gov/ticks/life_cycle_and_hosts.html

Hubálek, Z., Halouzka, J., Juricová, Z., Sikutová, S., & Rudolf, I. (2006). Effect of forest clearing on the abundance of Ixodes ricinus ticks and the prevalence of Borrelia burgdorferi s.l. *Medical and Veterinary Entomology*, *20*(2), 166–72. doi:10.1111/j.1365-2915.2006.00615.x

Jha, C. S., Goparaju, L., Tripathi, A., Gharai, B., Raghubanshi, A. S., & Singh, J. S. (2005). Forest fragmentation and its impact on species diversity: an analysis using remote sensing and GIS. *Biodiversity and Conservation*, *14*(7), 1681–1698. doi:10.1007/s10531-004-0695-y

Kalluri, S., Gilruth, P., Rogers, D., & Szczur, M. (2007). Surveillance of arthropod vector-borne infectious diseases using remote sensing techniques: a review. *PLoS Pathogens*, *3*(10), 1361–71. doi:10.1371/journal.ppat.0030116

Karimi, H. A. . (2014). *Big Data Techniques and Technologies in Geoinformatics* (pp. 177–192). CRC Press 2014. doi:10.1201/b16524-10

KNMI. (2015). Retrieved January 06, 2015, from http://www.knmi.nl/over_het_knmi/

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480. doi:10.1109/5.58325

Kohonen, T. (2001). The Basic SOM. In *Self-Organizing Maps* (pp. 105–176). Springer Berlin Heidelberg. doi:10.1007/978-3-642-56927-2_3

Kohonen, T. (2013a). Essentials of the self-organizing map. *Neural Networks : The Official Journal of the International Neural Network Society*, *37*, 52–65. doi:10.1016/j.neunet.2012.09.018

Laaksonen, J., & Honkela, T. (Eds.). (2011). *Advances in Self-Organizing Maps* (Vol. 6731, pp. 141–150). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-21566-7

Large-scale study of preventive antibiotic usage against Lyme disease - Wageningen UR. (2013). Retrieved August 10, 2014, from https://www.wageningenur.nl/en/show/Largescale-study-of-preventive-antibiotic-usage-against-Lyme-disease.htm?wmstepid=mail_de_auteur

Li, S., Heyman, P., Cochez, C., Simons, L., & Vanwambeke, S. O. (2012). A multi-level analysis of the relationship between environmental factors and questing Ixodes ricinus dynamics in Belgium. *Parasites & Vectors*, *5*, 149. doi:10.1186/1756-3305-5-149

Life cycle of Hard Ticks that Spread Disease. (2014a). Retrieved August 24, 2014, from http://www.cdc.gov/ticks/life_cycle_and_hosts.html

Life cycle of Hard Ticks that Spread Disease. (2014b). Retrieved September 08, 2014, from http://www.cdc.gov/ticks/life_cycle_and_hosts.html

Mulder, S., van Vliet, A. J. H., Bron, W. A., Gassner, F., & Takken, W. (2013). High risk of tick bites in Dutch gardens. *Vector Borne and Zoonotic Diseases (Larchmont, N.Y.)*, *13*(12), 865–71. doi:10.1089/vbz.2012.1194

Park, Y.-S., Kwon, Y.-S., Hwang, S.-J., & Park, S. (2014). Characterizing effects of landscape and morphometric factors on water quality of reservoirs using a self-organizing map. *Environmental Modelling & Software*, *55*, 214–221. doi:10.1016/j.envsoft.2014.01.031

Parola, P., & Raoult, D. (2001). Ticks and tickborne bacterial diseases in humans: an emerging infectious threat. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *32*(6), 897–928. doi:10.1086/319347

Petney, T. N., Skuballa, J., Muders, S., Pfäffle, M., Zetlmeisl, C., & Oehme, R. (2012). *Arthropods as Vectors of Emerging Diseases*. (H. Mehlhorn, Ed.) (Vol. 3, pp. 151–166). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-28842-5

Pfäffle, M., Littwin, N., Muders, S. V, & Petney, T. N. (2013). The ecology of tick-borne diseases. *International Journal for Parasitology*, *43*(12-13), 1059–77. doi:10.1016/j.ijpara.2013.06.009

Randolph, S. E. (2013). Ecology of non-nidicolous ticks. In R. M. R. Daniel E. Sonenshine (Ed.), *Biology of ticks* (2nd ed., p. 7). New York: Oxford University Press. Retrieved from http://books.google.nl/books

Rao, K. V., A. Govardhan, & Rao, K. V. C. (2012). Spatiotemporal Data Mining: Issues, Tasks And Applications. *International Journal of Computer Science & Engineering Survey*, *3*(1), 39–52. doi:10.5121/ijcses.2012.3104

Reye, A. L., Hübschen, J. M., Sausy, A., & Muller, C. P. (2010). Prevalence and seasonality of tick-borne pathogens in questing Ixodes ricinus ticks from Luxembourg. *Applied and Environmental Microbiology*, *76*(9), 2923–31. doi:10.1128/AEM.03061-09

Rosà, R., & Pugliese, A. (2007). Effects of tick population dynamics and host densities on the persistence of tick-borne infections. *Mathematical Biosciences*, *208*(1), 216–40. doi:10.1016/j.mbs.2006.10.002

Sandberg, S., Awerbuch, T. E., & Spielman, A. (1992). A comprehensive multiple matrix model representing the life cycle of the tick that transmits agent of lyme disease. *Journal of Theoretical Biology*, *157*(2), 203–220. doi:10.1016/S0022-5193(05)80621-6

Silva, B., & Marques, N. (2010). Feature Clustering with Self-organizing Maps and an Application to Financial Time-series for Portfolio Selection. In J. Filipe & J. Kacprzyk (Eds.), *IJCCI ICFC-ICNC* (p. page 301–309). SciTePress. Retrieved from http://ssdi.di.fct.unl.pt/~nmm/MyPapers/SM2010b.pdf

Sobrino, R., Millán, J., Oleaga, A., Gortázar, C., de la Fuente, J., & Ruiz-Fons, F. (2012). Ecological preferences of exophilic and endophilic ticks (Acari: Ixodidae) parasitizing wild carnivores in the Iberian Peninsula. *Veterinary Parasitology*, *184*(2-4), 248–57. doi:10.1016/j.vetpar.2011.09.003

Sprong, H., Hofhuis, A., Gassner, F., Takken, W., Jacobs, F., van Vliet, A. J. H., … Takumi, K. (2012). Circumstantial evidence for an increase in the total number and activity of Borrelia-infected Ixodes ricinus in the Netherlands. *Parasites & Vectors*, *5*, 294. doi:10.1186/1756-3305-5-294

Sprong, H., Trentelman, J., Seemann, I., Grubhoffer, L., Rego, R. O. M., Hajdušek, O., … Hovius, J. W. R. (2014). ANTIDotE: anti-tick vaccines to prevent tick-borne diseases in Europe. *Parasites & Vectors*, *7*(1), 77. doi:10.1186/1756-3305-7-77

Stojanova, D., Ceci, M., Appice, A., Malerba, D., & Džeroski, S. (2013). Dealing with spatial autocorrelation when learning predictive clustering trees. *Ecological Informatics*, *13*, 22–39. doi:10.1016/j.ecoinf.2012.10.006

Subak, S. (2003). Effects of Climate on Variability in Lyme Disease Incidence in the Northeastern United States. *American Journal of Epidemiology*, *157*(6), 531–538. doi:10.1093/aje/kwg014

Tack, W., Madder, M., Baeten, L., Vanhellemont, M., Gruwez, R., & Verheyen, K. (2012). Local habitat and landscape affect Ixodes ricinus tick abundances in forests on poor, sandy soils. *Forest Ecology and Management*, *265*, 30–36. doi:10.1016/j.foreco.2011.10.028

Tekenradar.nl (Tick radar). (2014). Retrieved August 25, 2014, from https://www.wageningenur.nl/en/show/Tekenradar.nl-Tick-radar.htm

Tijsse-Klasen, E., Fonville, M., Reimerink, J. H., Spitzen-van der Sluijs, A., & Sprong, H. (2010). Role of sand lizards in the ecology of Lyme and other tick-borne diseases in the Netherlands. *Parasites & Vectors*, *3*, 42. doi:10.1186/1756-3305-3-42

Tisu, M. (2012). *Spatio-temporal data mining for vineyard yeild estimation: A Slovenian case study and self-organizing maps*. Utrecht university. Retrieved from http://dspace.library.uu.nl/handle/1874/259141

Tran, P. M., & Waller, L. (2013). Effects of landscape fragmentation and climate on Lyme disease incidence in the northeastern United States. *EcoHealth*, *10*(4), 394–404. doi:10.1007/s10393-013-0890-y

Vesanto, J., & Ahola, J. (1999). Hunting for Correlations in Data Using the Self-Organizing Map. In *In Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications (CIMA'99), ICSC Academic Press* (pp. 279–285).

Vesanto, J., & Sulkava, M. (2002). Distance matrix based clustering of the Self- Organizing Map. In J. R. (Ed.), *12th International Conference on Artificial Neural Networks (ICANN 2002).* (pp. 951–956). Madrid, Spain,. Retrieved from http://lib.tkk.fi/Diss/2008/isbn9789512291540/article1.pdf

Vu Hai, V., Almeras, L., Socolovschi, C., Raoult, D., Parola, P., & Pagès, F. (2014). Monitoring human tick-borne disease risk and tick bite exposure in Europe: Available tools and promising future methods. *Ticks and Tick-Borne Diseases*. doi:10.1016/j.ttbdis.2014.07.022

Wiersma, Y. F. (2010). Birding 2.0: Citizen Science and Effective Monitoring in the Web 2.0 World. *Avian Conservation and Ecology*, *5*(2), art13. doi:10.5751/ACE-00427-050213

Wimberly, M. C., Baer, A. D., & Yabsley, M. J. (2008). Enhanced spatial models for predicting the geographic distributions of tick-borne pathogens. *International Journal of Health Geographics*, *7*, 15. doi:10.1186/1476-072X-7-15

World Health Organization. (2014). *WHO, A global brief on vector-borne diseases*. Retrieved from http://www.who.int/campaigns/world-health-day/2014/global-brief/en/

Yin, H. (2008). The Self-Organizing Maps: Background, Theories, Extensions and Applications. In P. L. C. J. Fulcher, Prof. John (Ed.), *Computational Intelligence: A Compendium* (115th ed., pp. 715–762). doi:10.1007/978-3-540-78293-3_17

Zhou, L., Wei, J., & Zhao, D. (2014). Detecting the Impacts of Socioeconomic Factors on Regional Severity of Work-Related Casualties in China. *Human and Ecological Risk Assessment: An International Journal*, *20*(6), 1469–1490. doi:10.1080/10807039.2014.892361

APPENDICES

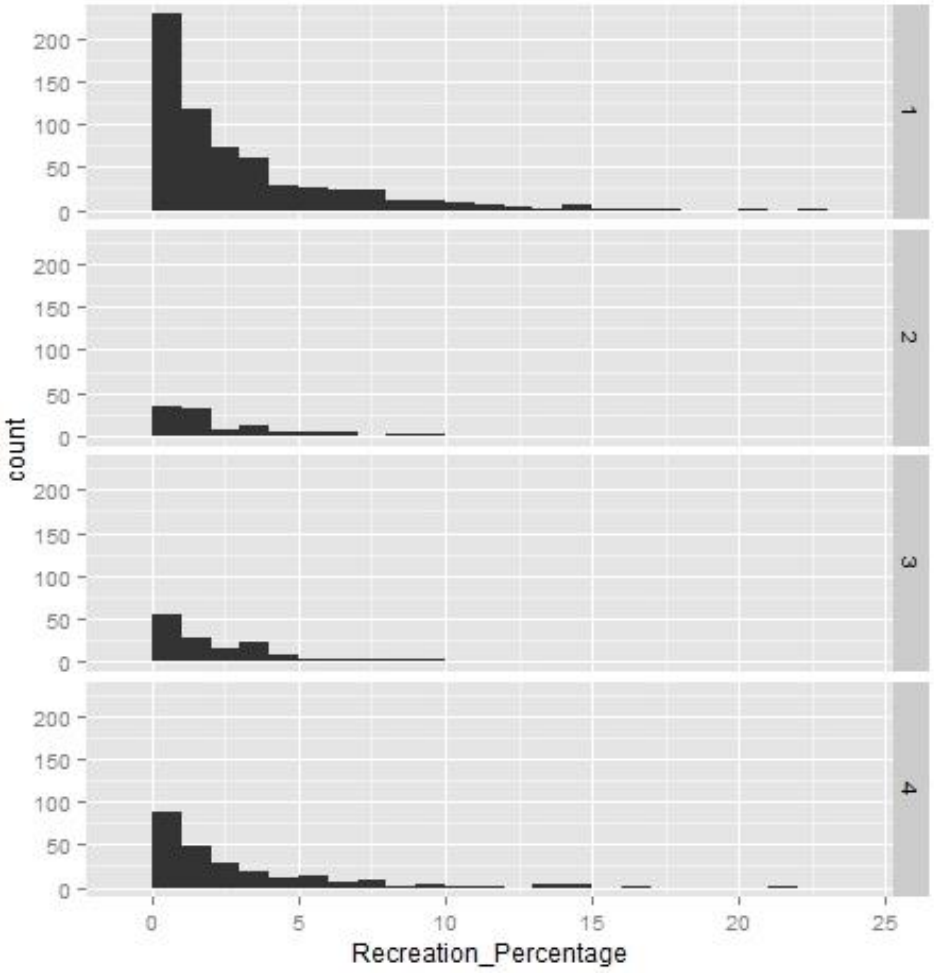
Appendix (A):

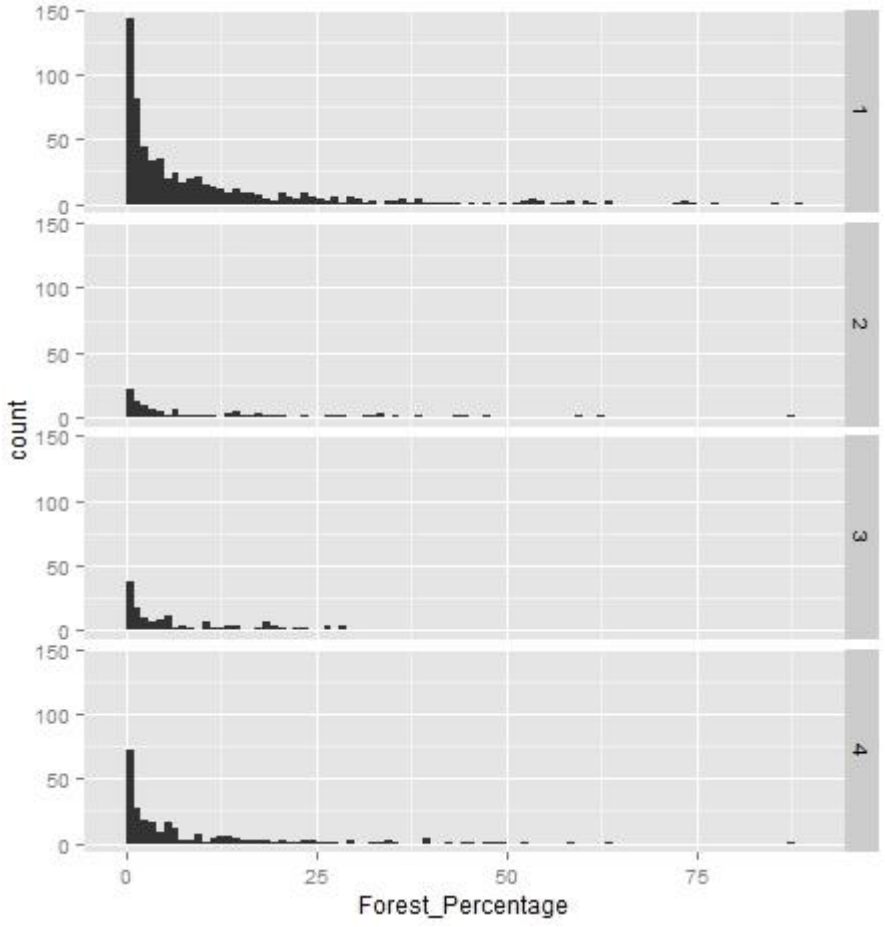
The result of Moran's I test.

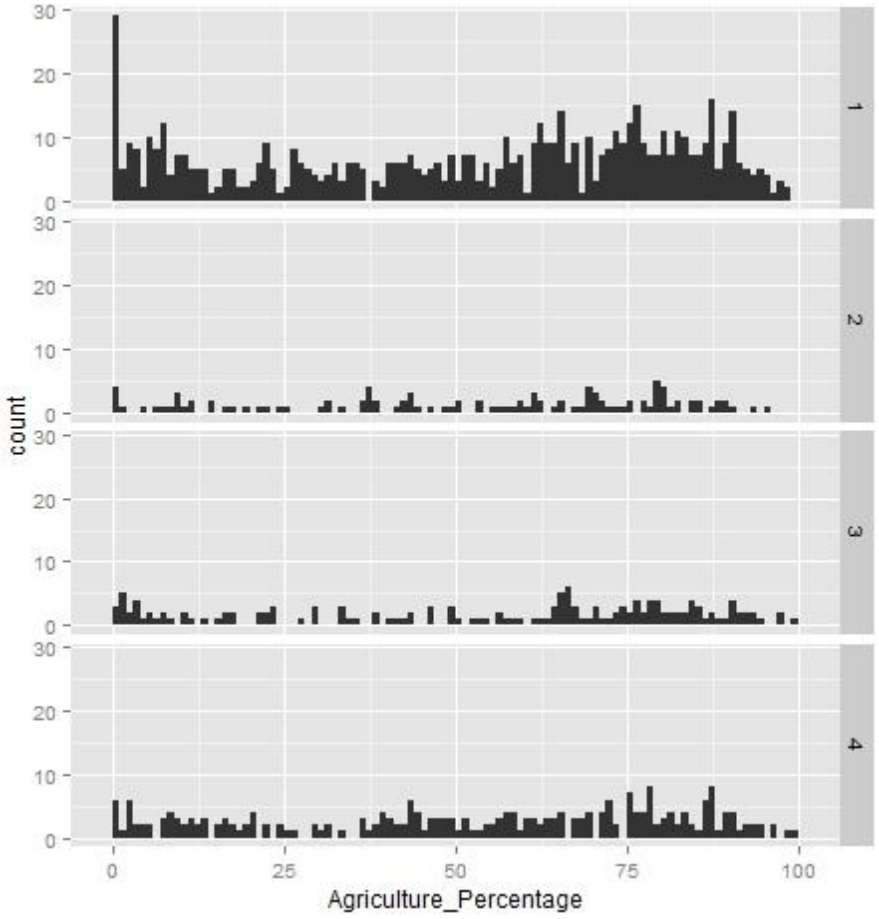
Year	Moran's I Index	z-score	p-value
2006	0.016	1.228	0.219
2007	0.034	2.894	0.003
2008	0.038	2.325	0.020
2009	0.039	3.422	<0.001
2010	0.022	1.907	0.056
2011	0.008	0.722	0.470
2012	0.273	23.628	<0.001
2013	0.371	25.889	<0.001
2014	0.139	9.697	<0.001

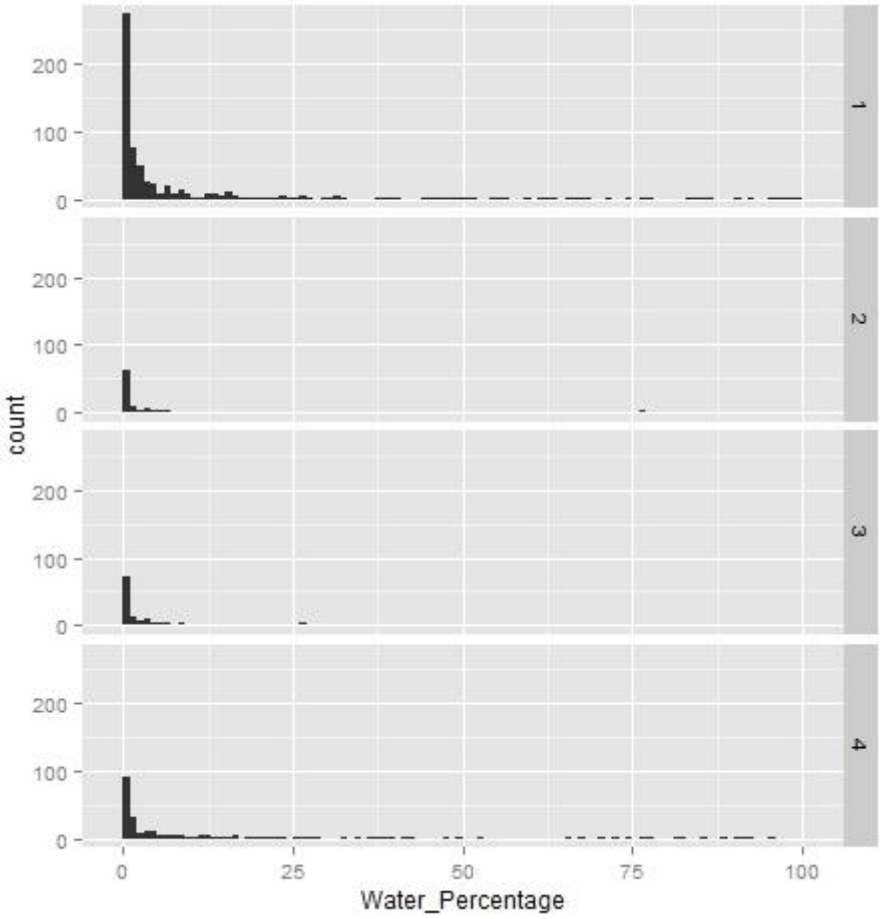
Appendix (B):

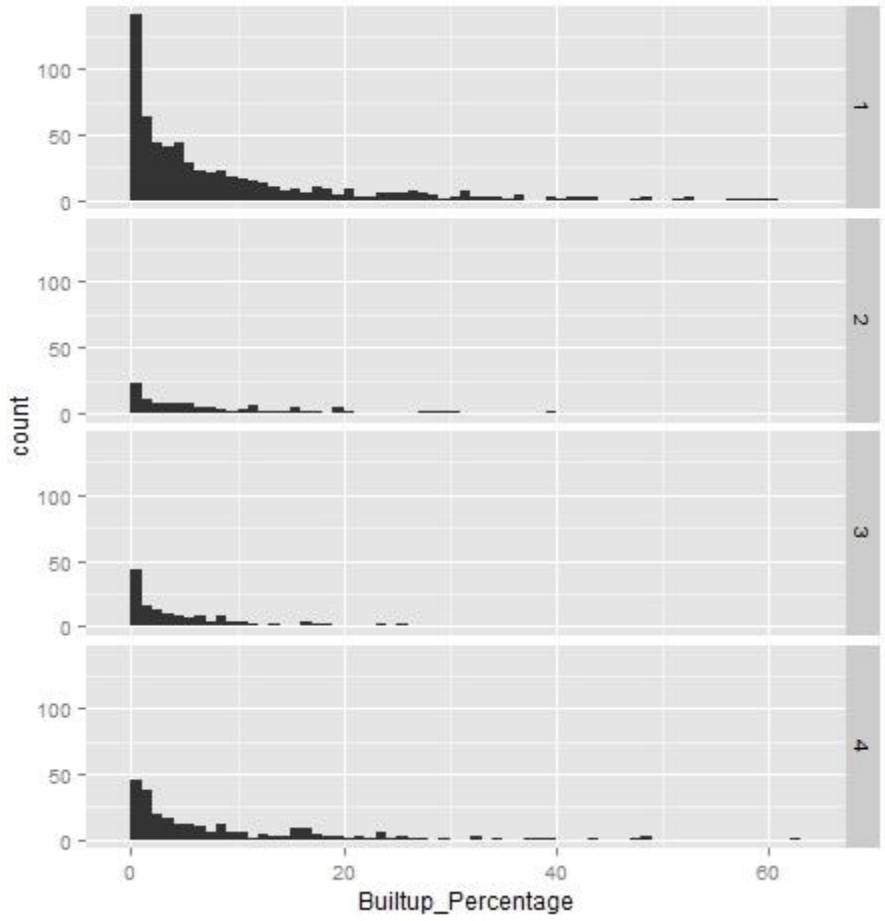
Histograms show the percentage of each land use land cover in the spatial units per cluster.




count
Forest_Percentage
0
25
50
75
150
100
50
1
2
3
4


count
Agriculture_Percentage
0
25
50
75
100
30
20
10
1
2
3
4


count
200
100
0
1
2
3
4
0
25
50
75
100
Water_Percentage


count
100
50
0
1
2
3
4
0
20
40
60
Builtup_Percentage

Appendix (C):

Data preparation code:

- To generate WOY from the date of the tick bite observations
- Create fishnet to divide the study area to regular size spatial units
- Spatial aggregation by counting the number of observations in each spatial units
- Calculate the LULC percentage in each spatial unit

```
import os
import csv
import datetime
import dateutil

os.chdir(r'D:\thesis\Data\newtrial')

f= open('all.csv','r')
reader=csv.reader(f,delimiter=';')
with open('all_WOY.csv', 'w') as csvfile:
    spamwriter = csv.writer(csvfile, delimiter=';')
    for line in reader:
        TickDate = line[0]
        split = TickDate.split("-")
        week = datetime.date(int(split[0]), int(split[1]), int(split[2])).isocalendar()[1]
        line.append(week)
        spamwriter.writerow(line)
    f.close()

import arcpy
from arcpy import env

# set workspace environment
env.workspace = "D:/thesis/Data/Required-data"
env.overwriteOutput=True
# to create fishnet ############################################################

# Set coordinate system of the output fishnet
env.outputCoordinateSystem = arcpy.SpatialReference("D:/thesis/Data/Required-data/NewGrid_label.prj")

# Set the origin of the fishnet
originCoordinate = '10425.20001220705 306846.200012207'
# Set the orientation
yAxisCoordinate = '10425.20001220705 306856.200012207'

# Enter the cell size:
cellSizeWidth = '5000'
cellSizeHeight = '5000'

oppositeCoorner = '278026.1000366211 621876.299987793'

numRows =  '0'
numColumns = '0'
# Create a point label feature class
labels = 'LABELS'

# Extent is set by origin and opposite corner
templateExtent = '#'

# Each  cell will be a polygon
geometryType = 'POLYGON'
outFeatureClass = 'D:/thesis/Data/Required-data/fishnet5by5.shp'
arcpy.CreateFishnet_management(outFeatureClass, originCoordinate, yAxisCoordinate, cellSizeWidth,
cellSizeHeight, numRows, numColumns, oppositeCoorner, labels, templateExtent, geometryType)
#del outFeatureClass
```

```
# to count points in polygon :#############################################################

outFeatureClass = 'D:/thesis/Data/Required-data/fishnet5by5.shp'
POINTtick = 'D:/thesis/Data/Required-data/code2/allyearsCopy.shp' # this file has tickbite reports
arcpy.MakeFeatureLayer_management(POINTtick,"point")

arcpy.MakeFeatureLayer_management(outFeatureClass,"polygon") # this represent the grid which created (fishnet)

#create update cursor

arcpy.AddField_management("polygon", "count", "LONG", 9, "", "","numpoint", "NULLABLE")
POLYGONgrid= arcpy.UpdateCursor("polygon")

try:
    for row in POLYGONgrid:

        print row.getValue("count"),

        # Select each record inside of the polygon feature class

        expression = "\"FID\" = " + str(row.getValue("FID")) + ""

        SelPoly = arcpy.SelectLayerByAttribute_management("polygon", "NEW_SELECTION",expression)

  # Select all the point that are inside of the polygon record
        SelPts = arcpy.SelectLayerByLocation_management("point", "WITHIN", SelPoly, 0, "NEW_SELECTION")

        # Count the points that are in each polygon
        GetCount = arcpy.GetCount_management(SelPts)
        GetCount = int(GetCount.getOutput(0))
            print expression + "  " + str(GetCount)
            row.setValue("count", GetCount)
            POLYGONgrid.updateRow(row)

        del POLYGONgrid, row
  except Exception as e:
        print e
        print (arcpy.GetMessages())


        # to calculate the area of each cell :#############################################################

        InputTable = "D:/thesis/Data/Required-data/fishnet5by5.shp"

        arcpy.AddField_management(InputTable,"cell_Area","DOUBLE","#","#","#","#","NULLABLE","NON_REQUIRED","#")
        arcpy.CalculateField_management(InputTable,"cell_Area","!shape.area@squaremeters!","PYTHON_9.3","#")

        # to caculate the percentage of each LULC type :#############################################

        grid = 'D:/thesis/Data/Required-data/fishnet5by5.shp' # this is fishnet file
        landuse='D:/thesis/Data/Required-data/New_results/ReclassifyBBG2008/Water.shp' # this is  LULC file
        outFeatureClass = 'Waterbodyfeature.shp'


        #calculate percentage of forest ###################################################################

        arcpy.Intersect_analysis([landuse,grid], outFeatureClass, "ALL", "")
        arcpy.AddField_management(outFeatureClass, "FArea_m2", "FLOAT", "", "", "", "", "NON_NULLABLE", "NON_REQUIRED", "")

        # Process: Calculate Field
        arcpy.CalculateField_management(outFeatureClass, "FArea_m2", "!shape.area@squaremeters!", "PYTHON_9.3", "")

        # Process: Add Field - percent of grid cell
        arcpy.AddField_management(outFeatureClass, "W_Percent", "FLOAT", "", "", "", "", "NON_NULLABLE", "NON_REQUIRED", "")


        arcpy.CalculateField_management(outFeatureClass, "W_Percent", "!FArea_m2! /25000000","PYTHON_9.3","")
```

Appendix (D):

Code to construct the similarity measure

```
import arcpy
from arcpy import env
import numpy as np


# set workspace environment
env.workspace = "D:/thesis/Data/Required-data/New_results/Temporal_Aggregation"
env.overwriteOutput=True
env.outputCoordinateSystem = arcpy.SpatialReference("D:/thesis/Data/Required-data/seperateyear_withoutwater/allyears.prj")

outFeatureClass = 'D:/thesis/Data/clip2013.shp' # Polygon shapefile in which the analyst wants to count the number of points.
POINTtick = 'D:/thesis/Data/Yearly_shapefiles/2013.shp' # Point shapefile (ticks bites reports) observations

arcpy.MakeFeatureLayer_management(POINTtick,"point")
arcpy.MakeFeatureLayer_management(outFeatureClass,"polygon")

#create update cursor

arcpy.AddField_management("polygon", "count", "LONG", 9, "", "","numpoint", "NULLABLE")
for week in range(52):
    arcpy.AddField_management("polygon", "week" + str(week + 1), "LONG", 9, "", "","week", "NULLABLE")
for cweek in range(52):
    arcpy.AddField_management("polygon", "cweek" + str(cweek + 1), "LONG", 9, "", "","cweek", "NULLABLE")

arcpy.AddField_management("polygon", "ind25", "LONG", 9, "", "","per25", "NULLABLE")
arcpy.AddField_management("polygon", "ind50", "LONG", 9, "", "","per50", "NULLABLE")
arcpy.AddField_management("polygon", "ind75", "LONG", 9, "", "","per75", "NULLABLE")
arcpy.AddField_management("polygon", "ind100", "LONG", 9, "", "","per100", "NULLABLE")
POLYGONgrid= arcpy.UpdateCursor("polygon")
allTickPoint = arcpy.SearchCursor("point")

allTickPoint = arcpy.SearchCursor("point")

try:
    for row in POLYGONgrid:

        # Select each record inside of the polygon feature class

        expression = "\"FID\" = " + str(row.getValue("FID")) + ""

        SelPoly = arcpy.SelectLayerByAttribute_management("polygon", "NEW_SELECTION",expression)

  # Select all the point that are inside of the polygon record

        SelPts = arcpy.SelectLayerByLocation_management("point", "WITHIN", SelPoly, 0, "NEW_SELECTION")

        arcpy.MakeFeatureLayer_management(SelPts, "selectedPoints")
        selPtsCursor = arcpy.SearchCursor("selectedPoints")

  # Aggregate the point obserations  weekly
        weekCounts = [0] * 52
        for point in selPtsCursor:
            for week in range(52):
                if point.getValue("Field13") == week + 1:
                    weekCounts[week] += 1
                    break

        for week in range(52):
            row.setValue("week" + str(week + 1), weekCounts[week])

  # Calculate cummulative sum from week 1 to week 52
        cumSum = [0] * 52
        for week in range(52):
            cumSum[week] = sum(weekCounts[0:week+1])
            row.setValue("cweek" + str(week + 1), cumSum[week])
```

```
    # Construct the similarity measure:
    # by Calculating the week number when 25% 50% 75% 100% of the cummulative sum reached.

          cumSum_np = np.array(cumSum)


          ind25 = np.argmin(abs(cumSum_np-(0.25*cumSum_np[51])))+1
          ind50 = np.argmin(abs(cumSum_np-(0.5*cumSum_np[51])))+1
          ind75 = np.argmin(abs(cumSum_np-(0.75*cumSum_np[51])))+1
          ind100 = np.argmin(abs(cumSum_np-(1*cumSum_np[51])))+1

          row.setValue("ind25", ind25)
          row.setValue("ind50", ind50)
          row.setValue("ind75", ind75)
          row.setValue("ind100", ind100)

          # Count the points in each polygon(spatial unit)
          GetCount = arcpy.GetCount_management(SelPts)
          GetCount = int(GetCount.getOutput(0))
          row.setValue("count", GetCount)
          POLYGONgrid.updateRow(row)


      del POLYGONgrid, row
  except Exception as e:
   print (e)
   print (arcpy.GetMessages())
```

Appendix (E):

Results of projecting SOM into the geographic space, Here, SOMs were trained based on several aggregation levels. Figure (A) :10km×10km , Figure (B) : 30km×30km , Figures (C and D): SOM trained based on Tekenradar dataset only for year 2012,2013.